TDT FOR HUMAN QTL MAPPING AND GENOME-WIDE ASSOCIATION STUDY

HAO YING

NATIONAL UNIVERSITY OF SINGAPORE

2008

TDT FOR HUMAN QTL MAPPING AND GENOME-WIDE ASSOCIATION STUDY

HAO YING

(Master of Science, National University of Singapore)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2008

Acknowledgements

This thesis would not have been possible without the support and help of many people. It is pleasant that I have now the opportunity to express my gratitude for all of them.

The first people I would like to thank is my supervisor, Associate Professor CHEN Zehua. He is a sympathetic and considerate person with much of enthusiasm and integral view on research. During the past four years, I am fortunate to receive his continuous support and learn a lot from him, not only the way to do research, but also the careful and precise manner to the scientific research. His patience and encouragement help me to overcome a lot of difficulties.

I am also grateful to Associate Professor Zhang Louxin. When I studied for my master degree at Mathematics Department, he provided me a lot of help in learning bioinformatics.

My gratitude also goes to the National University of Singapore awarding me a research scholarship, and the Department of Statistics and Applied Probability for providing the excellent research environment. During my Ph.D. programme I received con-

tinuous help from staffs at our department, especially our nice IT support Ms Yvonne and Dr. Zhang Rong who gave me much of help when I was running my programme.

I would like to thank my friendly colleagues Dr. Zhao Yudong for much of help on learning some computer softwares, and Dr. Li Wenyun and Dr. Liu Huixia for useful discussion with them during my study.

I feel a deep sense of gratitude for my husband Fanwen, for his love, thoughtfulness and patience during the PhD period and my lovely daughter Yuan-yuan for her accompanying and a lot of joy she gave me and hope that these efforts will inspire the same spirit in her .

Finally, I am greatly indebted to my parents and my auntie Mdm Hao Bingxin who never failed to encourage me and to support me whenever they could.

Contents

Summary

To find the genetic variants contributing to a complex disease, the researchers developed a lot of genetic mapping techniques. Association study and linkage study are the two main approaches. It is well-known that the possible stratification of the population can make us draw spurious association from the conventional case-control study. Spielman et. al. (1993) proposed a very efficient test, namely, transmissiom/disequilibrium test (TDT), which provides a valid test of linkage and association. TDT is intend to test the linkage between a genetic marker locus and disease causal locus by comparing the marker allele transmission times between the affected and the unaffected. A specificity of the TDT is that it is robust to the presence of population structure/stratification. As we know the population structure/stratification can cause the spurious linkage even when there is no genetic association between a marker locus and a trait locus. The TDT has attracted much interest in gene mapping for a complex disease and quantitative traits. Various TDT for quantitative trait locus (QTL) mapping have been developed. In this thesis, we will do our contributions in TDT in the following aspect.

To begin with, the sample size required might be too large for the application of TDT

in practice. Thus, it is very important to develop sampling schemes that can be carried out easily and reduce the cost of sampling. In this thesis, we provide a simple and efficient sampling approach for application of TDT in QTL mapping. The properties of this sampling scheme and the effect of selective genotyping on the power of TDT are studied. Simulation studies are also carried out to demonstrate the desirable power compared with conventional truncation sampling approach.

Furthermore, though the TDT approach used for gene mapping at multi loci has been studied by a number of researchers recently, the application of TDT to genome-wide association study has not been tackled so far. Since the rapid improvement in SNP genotyping technology makes it possible to find the genetic contributions to common disease, in this thesis we develop a generalized TDT by a penalized logistic model to extend the TDT to genome-wide association study. By virtue of this model, we convert the linkage study for gene mapping to variable selection problem. A two-step method which combines the efficient algorithm for variables selection with a new criterion for model selection is proposed. In the simulation study, by comparing the false discovery rate and positive selection rate with the Bonferroni-type multiple-comparison approach, it is demonstrated that our method is valid and efficient.

Finally, as genome-wide association study always gives us a model space of the large dimension, where a variable of interest is influenced by a number of potential co-variates, the issue of variable selection or model selection is very necessary to statistical data analysis with large dimension. In this thesis, in the generalization of TDT to the

genome-wide association study, a logistic model with grouped variables and the penalized likelihood are constructed. We study the optimality conditions for the maximizing the penalized likelihood and then provide a simple optimality criterion. According to this criterion, we further propose an efficient algorithm for variable selection in logistic model with grouped variables.

List of Tables

List of Figures

Chapter 1

Introduction

As we known, many human traits or disease are viewed as having some genetic component. In recent years, genetic study of human traits has received a great deal of attention, people have made a lot of effort to find and study the genes which involved in complex human traits. As a result, many statistical methods are continually being proposed and developed. In this chapter, we first summarize the general knowledge of molecular genetics and then briefly review some genetic mapping methods for human traits.

1.1 Genetics background

There are 23 pairs of *chromosomes* in human genome. Two of them are sex-chromosomes and the remaining 22 homologous pairs are termed *autosomal chromosomes*. Along a chromosome, at definite sites termed genetic loci, there are are *genes*. A gene consists

of several variants, namely *alleles*. For an individual, the pair of alleles (maternal and paternal) at a locus is called the *genotype*. If the two alleles are identical, then the genotype is called *homozygous*; otherwise, *heterozygous*. Along a single chromosome, the pattern of a sequence of alleles is called a *haplotype*, the two haplotypes for an individual is still called a (multilocus) genotype. At each locus, only one of two alleles is transmitted from a given parent to his/her offsprings, the *transmission* probability of each allele is 1/2.

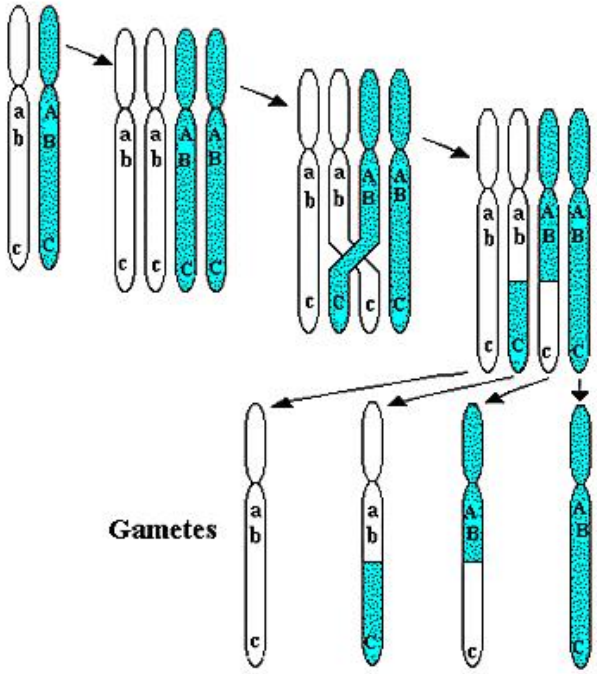
Generally, genotypes are unobservable without complicated biological experiments, what is observable is a person's *phenotype*. The phenotype may be discrete or continuous. The *penetrance* is the probability of a phenotype given a genotype. The genotype frequencies in a population can be calculated from the allele frequencies if the population is in *Hardy-Weinberg Equilibrium*. In particular, the frequency of a genotype is the product of the frequencies of the two constituent alleles.

In *Mendelian* law of inheritance, for simple Mendelian traits, the law of independence states that the genes are transmitted from parents to an offspring independently of one another. For example, if a person has blood group A (e.g. genotype AO) and brown eyes (e.g. genotype Bb, allele B is for brown and allele b is for blue eyes), the transmission of the alleles A and O to offspring is independent of alleles B and b. However, not all genes are transmitted independently of one another. Genes at two physically close loci on the same chromosome tend to transmit together. And furthermore, when several genetic loci on the same chromosome are simultaneously followed in a human

pedigree, the phenomenon of *recombination* can often be observed. That is, for a pair of chromosomes, neither the maternal nor paternal chromosome is completely transmitted, instead, a combination of the paternal and the maternal chromosome is transmitted. Recombination takes place during a process known as *meiosis* which is the type of cell division of producing gametes(egg or sperm), portions of paternal and maternal chromosomes interchange by *crossing over*.

This process is illustrated in Figure 1.1. If an odd number of crossovers occur between two loci, there is a recombination between the two loci. The closer the loci are together, the smaller the probability of a recombination. The *recombination fraction* θ ($0 \leq \theta \leq 1/2$) between two loci is the probability that a recombination occurs between them. If the two loci are far apart, allele at one locus is transmitted independently of allele at the other, and $\theta = 1/2$, the two loci are said to be *unlinked* if θ is less than 1/2 they are said to be *linked*. The *genetic distance* is measured by *Morgan (M) or centiMorgan (cM),* $1M = 100cM$. The distance between two loci is 1 Morgan if the expected number of crossovers between the two loci is 1.

Figure 1.1: Crossing over and recombination in the process of meiosis



In the formation of gametes, the precursor cells of the sperm or ova double and form two chromosomes of a homologous. When the two chromosomes pair off in cell division in meiosis, they may exchange segments in the manner shown above and produce genetic variations in germ cells (two gametes contain different combinations of genes with the precursor).

Linkage equilibrium refers to that the population frequency of a two-locus haplotype is the product of the frequencies of the single-locus alleles. Assume we have two diallelic loci, one a disease locus with alleles A and a (with frequencies p_A and $1 - p_A$ respectively), and the other a marker locus with alleles B and b (with frequencies p_B and $1 - p_B$ respectively),the follows show the four haplotype frequencies under linkage equilibrium:

$$\begin{aligned} h_{AB} &= p_A p_B \\ h_{Ab} &= p_A(1 - p_B) \\ h_{aB} &= (1 - p_A)p_B \\ h_{ab} &= (1 - p_A)(1 - p_B). \end{aligned}$$

There is a *linkage disequilibrium* between these two loci, if there is number D such that the haplotype frequencies are as follows:

$$\begin{aligned} h_{AB} &= p_A p_B + D \\ h_{Ab} &= p_A(1 - p_B) - D \\ h_{aB} &= (1 - p_A)p_B - D \\ h_{ab} &= (1 - p_A)(1 - p_B) + D. \end{aligned}$$

Here D is the departure from equilibrium, namely linkage disequilibrium coefficient. Linkage disequilibrium may be resulted by linkage between two loci, population admixture/stratification or genetic mutation or selection. Under random mating, the magnitude of disequilibrium dissipates over the time of generations and linkage equilibrium

is reached from linkage disequilibrium. After k generations, disequilibrium decays according to the formula

$$D_k = (1-\theta)^k D_0,$$

where θ is the recombination fraction. It is apparent that D depends on the frequency of the alleles inspected and D is maximal when the allele frequencies all are 0.5. Thus, a normalization of D is suggested by dividing it with the theoretical maximum for the observed allele frequencies:

$$\begin{aligned} D' &= \frac{D}{D_{max}}, \quad \text{when} \quad D \geq 0, \quad \text{and} \quad D_{max} = \min\{p_A p_b, p_a p_B\}; \\ D' &= \frac{D}{D_{min}}, \quad \text{when} \quad D < 0, \quad \text{and} \quad D_{min} = \max\{-p_A p_B, -p_a p_b\}. \end{aligned}$$

When we consider the genetic contribution to a phenotype, this phenotype can be described as a qualitative trait (e.g disease or non-disease) or a quantitative trait (e.g arterial blood pressure). The relationship between a quantitative trait x and the genotype of the quantitative trait locus (QTL) (assume there is unique locus which has genetic contribution to the specific quantitative trait) can be expressed as follows:

$$x = g + e$$

where g is the effect of the genotype of QTL, e is a random variable which represents the combined effect of all non-genetic factors, generally it comes from a standard statistical distribution (e.g normal distribution). The genetic effect is fixed, i.e. the people carrying the same genotypes have the same value of g. For example, for a biallelic locus (A and a denote the name of alleles), there will be three different genotypes, so g can take

values μ_0, μ_1 and μ_2 for genotypes *aa*, *Aa*, and *AA* respectively. According to the *Law of Hardy-Weinberg Equilibrium*, we can determine the distribution of quantitative trait x and calculate the mean and variance due to QTL.

Let p denote the allele frequency of A, so the allele frequency of a is $1 - p$. The overall population mean will be a weighted average of the genotypic mean:

$$\mu = p^2\mu_0 + 2p(1-p)\mu_1 + (1-p)^2\mu_2.$$

The genetic variance can be calculated as follows:

$$V_g = p^2(\mu_0 - \mu)^2 + 2p(1-p)(\mu_1 - \mu)^2 + (1-p)^2(\mu_2 - \mu)^2.$$

Let V_e denote the variance due to non-genetic factors, then the total variance of quantitative trait x is $V_g + V_e$. The *heritability* is defined as

$$h = \frac{V_g}{V_g + V_e}.$$

It is a measure of the relative importance of the QTL.

1.2 Methods of genetic mapping

A complex disease refers to a disease determined by multiple genetic and environmental factors, perhaps their interactions as well. One of the aims of genetic study is to determine the biological contributions to a certain disease, i.e, to determine the location of risk genes. Such study is referred to us disease gene mapping (or QTL mapping

if the phenotype of disease is quantitative). There are two major mapping techniques, linkage analysis and association analysis. Generally, linkage analysis is conducted by using family data. It allows us determine regions of chromosomes that are likely to contain a risk gene based on genetic markers on these chromosomes. *Genetic markers* are the regions of DNA that vary between individuals. Linkage studies are based on the recombination. In linkage studies, some markers and the disease-causing genes are assumed to be located near each other on the same chromosome so that they tend to be inherited together. Thus we aim to search for the marker that is consistently present in those with a certain disease, and is not present in those without the disease. When a marker is found with the presence of the disease, we are able to know that there is a disease-causing gene somewhere close to the marker.

In linkage analysis study, one of the traditional methods is LOD score method (Ott, 1991, Morton, 1998). Assume that we can completely specify the mode of inheritance of the trait being studied, that means we have known the number of loci involved, the number of alleles at each locus and their frequencies, and the penetrance of each genotype. For a pedigree data, LOD score measures the likelihood of genetic linkage between loci. It is the log (base 10) odds ratio of linkage (the marker locus and trait locus are linked) and non-linkage (the marker locus and trait locus are not linked). LOD score $Z(\theta)$ is defined as follows:

$$Z(\theta) = \log_{10} \frac{\text{joint probability of all genotypes with a given } \theta}{\text{joint probability of all genotypes with } \theta = 1/2}.$$

Conventionally, for a given $\theta = \hat{\theta}$), $Z(\hat{\theta}) > 3$ is taken as the criterion for accepting

linkage.

Without assuming any specific mode of inheritance, another general method of linkage analysis is based on studying the association between phenotype and IBD (identity by descent) sharing of a marker locus among the family members such as pairs of siblings. This approach have attracted much of interests. See, Risch and Zhang (1995), Kong and Cox (1997), Hauser and Boehnke (1998), Dudoit and Speed (2000), for instance.

After using linkage analysis to get an idea where risk genes may be located, we can use association studies to try to locate the risk gene. We will test candidate genes to see if they are associated with presence of the disease. These tests can result in the location of a risk gene. However, association studies do not use families. Rather, they compare genotypes of affected individuals to genotypes of non-affected individuals, and these individuals do not have to be relatives.

Linkage disequilibrium (LD) mapping is another important mechanism for identifying genes underlying diseases. As we have known, a disease related to genetic disorder was resulted by mutations or immigration of carriers of mutant alleles into a population. At the initial time of a mutation on a particular location of the chromosome, it was completely associated with the adjacent marker alleles. Also, this association would remain over considerably many generations. As consequence of recombination, markers in the immediate vicinity of the disease locus are more likely to remain in the same strand than those that are farther away. One can estimate whether a particular marker locus ap-

pears to be in disequilibrium with a disease locus. In particular, if specific marker allele frequencies are higher in affected subjects than in unaffected subjects, this may suggest that there is linkage disequilibrium between a marker allele and the disease allele thus linkage between marker locus and disease locus. Xiong and Guo (1997) pointed that it is possible to map genes at a scale finer than 1 cM by the identification of markers that are in strong linkage disequilibrium with the disease allele. For the data summarized in Table 1.1, chi-square test statistics of LD mapping can be written as follows:

$$T_{LD} = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

Table 1.1: Observed marker allele counts in affected and unaffected subjects

	Number of alleles		
	M	m	Total
Affected	a	b	$a + b$
Unaffected	c	d	$c + d$
Total	$a + c$	$b + d$	N

Table 1.2: Population Stratification in Case-control Study

	Population 1		Population 2	
Genotype	case	control	case	control
MM	4	20	1	5
Mm	4	8	2	10
mm	2	12	17	85
Total	10	40	20	100
M allele frequency is 0.6 in case			M allele frequency is 0.1 in case	
M allele frequency is 0.6 in control			M allele frequency is 0.1 in control	

In LD mapping, we need to compare the frequencies of a specific marker allele between affected and unaffected subjects. However, an apparent LD in a population genetics structure might be caused by admixture or heterogeneity of the population, i.e, the stratification in the population. These factors may lead to spurious linkage from a conventional case-control study. Sometimes we may come up with substantial association even for unlinked loci. A simple example in Table 1.2 demonstrates such situation. From Table 1.2, we can see there is no difference of allele frequency between case and control in both populations. But the difference is induced by the pooled population Table 1.3.

To solve such problem, a very efficient approach was developed by Spielman et al. (1993), which is called *transmission/disequilibrium test* (TDT). We have known that if there exits some association between the phenotype and the underlying locus, then

Table 1.3: Pooled Population

	case	control
MM	5	25
Mm	6	18
mm	19	97
Total	30	140
	M frequency is 0.27	M frequency is 0.49

it may be the result that the allelic variation causes the phenotype directly or through an intermediary phenotype, or due to linkage disequilibrium between marker locus and trait locus. Under the assumption that the alleles at this locus assort independently during gametogenesis and the alleles are transmitted to offspring with probability $1/2$, i.e., there is no segregation distortion, the association also can be found between children's phenotype and the transmission of the two alleles at the causal locus. Thus, TDT takes the parent-child trios in which one or more parent is heterozygous at the underlying locus and compares the frequencies of the alleles transmitted from heterozygous parent to offspring with those of the alleles that are not transmitted.

The original intended use of the TDT was to test for linkage with a marker located near a candidate gene, in the cases where disease association had already been found. However, even in the absence of prior association study, the TDT is still valid. In other words, the TDT provides a joint test of linkage and association. Therefore, the TDT has attracted much interest in identification of genes for complex diseases and quantitative

traits. The related discussion can be found in recent references such as Sham et al. (1995), Xiong et al. (1998), Koeleman et al. (2000), Betensky et al. (2000), Deng et al. (2001, 2002), Sebastiani et al. (2004), Zhao et al. (2007).

1.3 Original idea of transmission/disequilibrium test

In this section, we introduce the detailed process of TDT. The TDT is a powerful family-based rather than population-based method to locate disease genes. It compares the number of times that heterozygous parents transmit the putatively associated marker allele to an affected child with the number of times that they transmit the alternate marker allele. For simplicity, we consider only biallelic loci. Assume there is a disease locus with two alleles D and d, and D is the disease allele, and a marker allele with alleles M and m linked to disease allele. Assume there is linkage disequilibrium between allele D and M. We consider case-parent trios, that is, the families with single affected child and at least one heterozygous parent at the marker locus. Apparently, for such families, the haplotype-sharing tests are not suitable since they require affected sib pairs. Suppose that we have n such case-parent trios. Assume further that there are w heterozygous Mm parents in these families and each parent is described in terms of the transmission status of M allele or m allele. The data can be simply described as follows:

$i =$ number of parents who transmit M to the child;

$j =$ number of parents who transmit m to the child.

Without segregation distortion, each of the marker alleles should be transmitted from heterozygous parents to their offspring with equal probabilities unless there is linkage and linkage disequilibrium (LD) between marker locus and disease susceptibility loci. Thus, without segregation distortion, the expectation of the number of parents who transmit allele M to the child is $E(i) = w/2$ and the expectation of the number of parents who transmit allele m to the child is $E(j) = w/2$. Furthermore, when $\theta = 1/2$ (no linkage) the two heterozygous parent transmit the marker alleles independently. Therefore, by McNemar's test, to test the null hypothesis of no linkage or no LD between the marker locus and the disease susceptibility loci, the test statistic can take the form of χ^2 statistic

$$\chi^2_{TDT} = (i - j)^2/w.$$

Under the null hypothesis, χ^2_{TDT} follows a asymptotic χ^2 distribution with one degree of freedom.

When there are two affected children in a family, the above test statistic is still valid because the allele contributions of two heterozygous parents to any one affected child are independent under the null hypothesis $\theta = 1/2$. For w heterozygous parents, we can summarize the data as follows:

$$\begin{aligned} i &= \text{number of parents who transmit M to both children;} \\ j &= \text{number of parents who transmit m to both children;} \\ w - i - j &= \text{number of parents who transmit M to one child and m to the other.} \end{aligned}$$

Thus, $2i+w-i-j$ M alleles and $2j+w-i-j$ m alleles are transmitted from heterozygous parents. Analogously, the χ^2 statistic is

$$\chi^2_{TDT} = (2i - 2j)^2/2w = 2(i - j)^2/w.$$

In general, if we consider families with more than two affected children, for example, families with three affected children. Then, each of w heterozygous parents belongs to one of the following four categories:

i = number of parents who transmit M to all three children;

l = number of parents who transmit M to two children and transmit m to the other child;

k = number of parents who transmit M to one child and transmit m to the other two children;

j = number of parents who transmit m to all three children.

The analogous test statistic is

$$\chi^2_{TDT} = (3i + l - k - 3j)^2/3w.$$

For the real data set, there may be varying numbers of affected children in different families. The TDT statistic can be obtained by pooling all affected children altogether

thus the χ^2 statistic can take the form of

$$\frac{(n_M - n_m)^2}{n_M + n_m},$$

where n_M and n_m denote the total numbers of transmitted M alleles and m alleles from heterozygous parents, respectively.

We note that the above TDT statistics use the affected child trios only, this test is valid for testing linkage and LD under an implicit assumption that there is no segregation distortion at marker locus. To ensure that the difference found with the above TDT statistics are not due to segregation distortion, Spielman et al. (1993) suggested the unaffected children data to be used as well. Thus, we can use a standard 2×2 contingency table χ^2 test to compare frequency of transmission of M between affected and unaffected children. In such situation, the data set is given in Table1.4.

Table 1.4: Number of marker alleles transmitted

	Alleles transmitted		
	M	m	Total
affected	n_M	n_m	$n_M + n_m$
unaffected	u_M	u_m	$u_M + u_m$
Total	$n_M + u_M$	$n_m + u_m$	w

The TDT statistic which is applied to both case-parent and unrelated control-parent trios (with at least one parent being heterozygous) can be written as

$$\chi^2_{TDT} = \frac{w(n_M u_m - n_m u_M)^2}{(n_M + n_m)(u_M + u_m)(n_M + u_M)(n_m + u_m)}. \tag{1.1}$$

Deng and Chen (2001) studied the detailed comparison of this TDT with TDT applied to case-parent trios only. In most of cases, this TDT is more powerful than the latter.

The TDT can be generalized to a marker locus with multi alleles. One simple approach is to consider the particularly interested allele as M, and all other markers are grouped as m in the case where more than one alleles are suggested to associate with the disease. For example, suppose that there are three alleles, M_1, M_2 and M_3, at the marker locus and both M_1 and M_2 are putative disease alleles. In this case, we can test both relevant heterozygotes by the TDT. This means that we can compare the transmission of M_1 to affected child by M_1M_3 parents with the transmission of M_3. Meanwhile, we can compare the transmission of M_2 to affected child by M_2M_3 parents with the transmission of M_3 as well. In these both two cases, we can use the same TDT statistic for the case where there is only one child in the family.

1.4 Literature review

As one of the approaches of LD mapping, the TDT was originally described in human genetics to test for linkage between a genetic marker and a disease-susceptibility locus. This technique was also applied in experimental species. Bink et al. (2000) used TDT in pig selection experiment for mapping loci affecting quantitative traits (growth performance). The TDT was used for screening a disease-susceptibility locus when it was

developed by Spielman et al. (1993). After that, many researchers extended it to QTL mapping, see Rabinowitz (1997), Allison (1997), Xiong et al. (1998), Sun et al. (2000), Zhu et al. (2001), Abecasis et al. (2000), Gauderman (2004), for instance.

Specifically, Rabinowitz (1997) took a statistic which reflects association between the value of the quantitative traits and the transmission of the given allele while Allison (1997) extended the TDT by examining the difference between average phenotypic values of children with different alleles transmitted from the heterozygous parent. Note that Allison's test is restrictive in that it requires family trios consisting of one heterozygous parent, one homozygous parent and one child, and this method is applicable for the situation of complete linkage disequilibrium only.

Recently, without these restrictions, Xiong et al. (1998) developed a χ^2 TDT statistic which is more powerful than Allison's. The above extension of TDT to QTL mapping was applied to families with both parents available. Later, Sun et al. (2000) proposed a new class of tests based on the work of Rabinowitz (1997) and Xiong et. al. (1998), which is applicable to families with only one parent available. Additionally, Knapp (1999), Sebastiani et al. (2004), and Croiseau et al. (2007) also studied the TDT with missing data, i.e., using genotypes of affected individuals and only one available parent of the individuals.

Other than the TDT applied to the family trios, some researchers also studied the extension of TDT to pedigree data. The big difference between family trios and pedigree data is non-independence of observations. George et. al. (1999) proposed a two-

stage procedure to detect the population association first and then to detect the linkage. Abecasis et al. (2000) also used the strategy which treats linkage and association separately.

Further, there has been a number of other research studies on the test of association and linkage without the sampling restrictions, such as, Rabinowitz et al. (2000), Monks and Kaplan (2000), Abecasis et al. (2001), and Martin et al. (2001), for instance.

Basically, the original TDT approach was used to test linkage and linkage disequilibrium between a single bi-allelic marker locus and the disease-susceptibility locus, so it can be treated as an application of McNemar's test, where the matched pairs are the parental alleles. To make the extensions of TDT to multi-allele marker locus. Various researchers proposed different approaches, see Duffy (1995), Bickeboller et al. (1995), Schaid (1996), Cleves et al. (1997), Lazzeroni et al. (1998), to name but a few.

Among these extensions, it is known that the maximal TDT proposed by Schaid (1996) is a very efficient method. The maximal TDT statistic is the maximum of the bi-allelic TDT statistics computed for each allele versus all others combined. The power of this test is likely to be high when only one allele is associated with the disease. But the shortcoming of this maximal TDT is that statistic is not a χ^2 random variable. Betensky et al. (2000) proposed a refinement to Bonferroni's correction for multiple testing for calculation accurate upper bounds fr the type I error and p-values for maximal TDT.

For extensions of TDT to multi loci, Sham et al. (1995) suggested a method called ETDT, which is analogous to logistic regression. The ETDT is adapted to CETDT

which is developed by Koeleman et al. (2000). CETDT which is robust to Hardy-Weinberg disequilibrium in the parents, is a test for an effect at a secondary locus or marker conditioning on the association of a candidate disease locus in case-parent trios. Some other people also studied the extensions to the TDT for multi loci (e.g. Betensky et al. (2000), Dudbridge et al. (2000)). But these extensions are not applicable to genome-wide association study.

1.5 Aim and organization of the thesis

TDT as a test of linkage employing nuclear family was developed to control for population admixture/stratification. It is originally designed for complex disease and has recently been extended to quantitative traits. However, a large sample size is required to detect QTL effects. This is especially true for identifying the genes with relatively small effects. Therefore, it is of great practical importance to developing sampling schemes that can effectively enhance the power and reduce the sample sizes required for TDT. In this thesis, we provide a simple and efficient sampling approach for TDT to extend TDT to quantitative trait locus mapping. Compared with traditional truncation method, our sampling scheme is much easier to carry out in practice and more economical. The effect of this sampling scheme on the joint distribution of genotypes of nuclear family is also studied..

The TDT was first applied to the linkage study of a single marker locus with the

disease-susceptibility gene. Although some researchers have extended it to multi marker loci, TDT has not been applied to genome-wide association study yet. In this thesis, we provide an efficient logistic model and an algorithm to search the causal genes in the genome-wide scope. This algorithm consists of crude selection and refined selection. In the step of crude selection, the sparse solutions, i.e., the sets of markers which may contain possibly effective genes are obtained by solving a penalized likelihood. Then, in the following step of refined selection, a new criterion for variable selection, namely EBIC, is applied to sift the final set of genes from those obtained in the first step. In simulation studies, we compare the positive selection rate and false discovery rate of EBIC with the traditional BIC and traditional multi-comparison method. In this thesis, we also propose a new approach of gene selection for grouped variable via the logistic model.

The thesis is organized as follows.

In chapter 2, we first give some preliminary materials to introduce issues of model selection, then, we investigate the logistic model with grouped variables and establish a new penalized likelihood model using a mathematical programming approach, which is analogous to LASSO to some extent. Then, we explore the optimality conditions for the underlying optimization problem and propose a new algorithm for this grouped variable selection via logistic model.

In chapter 3, we derive the joint probability distribution that the parent of the nuclear family is heterozygous and the offspring has the disease allele. We prove that the new

sampling scheme-extreme rank sampling (ERS) can increase this probability. We apply ERS to three different TDT statistics and conduct the simulation study to compare the power of the TDT under our sampling approaches and the truncation approach. The power of different TDTs are compared as well.

In chapter 4, the generalized TDT via logistic model is formulated and a two-step method is proposed to search for genes in the whole genome. By simulation study, we compare our approach with the conventional multi-test scheme, i.e., Benferroni-type method.

In chapter 5, we conclude the thesis and provide some possible directions of further research.

Chapter 2

Preliminaries

Variable selection with large model space has drawn increasing considerable attention in recent years. People do not know for a certainty which covariates are related to the response. So they might consider as many features as possible in the regression although the number of causal features is small. Nevertheless, the sample size is usually small. This gives rise to the so-called *sparse small-n-big-p* problem; that is, the number of candidate features, p, is much larger than the sample size n. For example, with the development of experimental techniques in genetics study, the genome-wide association study becomes possible. People has been able to type and locate tens or hundreds of thousands single nucleotide polymorphisms (SNPs) over the whole human genome. But there are only handful of them that are responsible to the genetic variation of a quantitative trait or a disease status. Another example is in a microarray chip. The expression values of thousands of genes in the specific tissues are measured and analyzed

in a microarray to identify a few tens of genes which can used for the classification or diagnose of the disease. In genetic studies, a very simple method is as follows. By fitting an appropriate model to features one at a time, the features with highest significant effects are selected by multiple-comparison approach, namely Bonferroni adjusted threshold value to control the family-wise type I error rate or the false-discovery rate (FDR). Apart from this approach, many more advanced approaches have been developed in the past recent years.

In this chapter, we briefly review some general methods for variable selection, then propose a new algorithm for penalized logistic model with grouped variables in a special case.

2.1 Introduction of variable selection

Suppose $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$, are n independent and identically distributed random vectors, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the p-dimensional predictor variables and y_i is the response variable of interest. We would like to model the relationship between the predictors and the response variable. We assume the following typical linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \tag{2.1}$$

where β_0 and $\beta = (\beta_1, \cdots, \beta_p)$ is the intercept and the regression coefficients associated with p predictors respectively, e_i is an error term.

In many practical situations, we don't know exactly which factors are related to the response, so we consider as many covariates as we can, hence not all these p covariates in the regression model are related to the response, some of them are superfluous, that means, conditional on a subset of the covariates, the response does not depend on the other covariates. The problem of variable selection is to identify the set of important predictors among p variables. Many traditional variable selection methods have been developed, such as, forward selection procedure, backward elimination procedure and best subset selection. Forward selection procedure starts with an empty subset, variables are added at a time until the optimality of some chosen criterion is reached. While backward elimination starts with the full subset,variables are removed one at a time. In best subset regression, we compare models made up of all possible subsets of candidate covariates. The best subset is usually intractable for large quantity of candidates because we have 2^p such subsets. These methods are used with some chosen model selection criteria, such as, Akaike information criterion (AIC) proposed by Akaike (1973) which is defined as

$$\text{AIC} = -2\log(L(\hat{\theta})) + 2K,$$

where $L(\hat{\theta})$ is the numerical value of the log-likelihood at its maximum point, $\hat{\theta}$ is the maximum likelihood estimate of parameter θ and K is the number of parameters used in the model. In application, one computes AIC for each of the candidate models with the covariates selected by a certain variable selection approach, and the model with the smallest value of AIC is choosen. Another classical variable selection criterion is Bayes

Information Criterion (BIC, Schwarz, 1978). It is simply

$$\text{BIC} = -2\log(L(\hat{\theta})) + K\log(n),$$

where n is the sample size. The BIC is similar to the AIC but having a stronger penalty for large models, tends to select models with fewer predictors than the AIC.

Besides the above variable selection methods, there are many other methods such as cross-validation method (Stone, 1974), Mallows' Cp (Mallows, 1973). These methods are usually too liberal, that is, they tend to select a model with many spurious covariates. Apart from these methods, more advanced approaches have been developed in the past recent years. For example, one efficient and popular approach called least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996). Compared with traditional estimation methods, the key difference of LASSO is that it has the ability to carry out parameter estimation and variable selection simultaneously. Fu (1998) developed a general approach to solve for the bridge estimator. Fan et al. (1999) proposed a few new approaches to select variables for linear models, robust regression models and generalized linear model based on a penalized likelihood method. Fan and Li (2001) studied the penalized likelihood method in linear regression, of which the LASSO is a special case. Park et al. (2006) introduced a path-following algorithm for L_1 regularized linear model.

In the following sections we start with introducing LASSO technique developed by Tibshirani (1996) and an extension of LASSO to the regression with grouped variable (Yuan et al. 2006) in linear regression. Next, LARS model selection algorithm (Efron

et al 2004) and finally an method to solve the sparse logistic regression (Sheevade and Keerthi 2003) are introduced.

2.2 LASSO and group LASSO in linear regression

2.2.1 LASSO

LASSO is a L_1 norm penalized likelihood approach for linear regression problem. The estimates of the coefficients in (2.1) are minimizing the penalized sum of squares:

$$\min_{\beta} \quad \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\lambda \geq 0$. The above problem is equivalent to solving

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \\ \text{subject to} \quad & \sum_{j=1}^{p}|\beta_j| \leq t. \end{aligned} \tag{2.2}$$

Here $t \geq 0$ is a tuning parameter. It controls the amount of shrinkage that is applied to the estimates. Let $\hat{\beta}^0$ be the ordinary least squares estimates and let $t_0 = \sum_j |\hat{\beta}^0_j|$. If $t \geq t_0$, the LASSO estimates are the same as the ordinary least squares estimates $\hat{\beta}^0_j$. If $t \leq t_0$, the LASSO estimates will shrink towards to 0, and some coefficients may be

forced to 0. This property enables LASSO to have the function of variable selection. For example, if $t = (1/2)t_0$, roughly $p/2$ of the coefficients will be forced to zero and $p/2$ of the variables will be selected.

For the solution of the above optimization problem, Tibshirani (1996) suggested two iterative algorithms. The first algorithm treats it as a problem with $p+1$ variables subject to 2^p constraints. The constraint $\sum_{j=1}^{p} |\beta_j| \leq t$ is equivalent to 2^p constraints which are of the form $\delta_{\mathbf{i}}^{\mathbf{T}}\beta \leq t, i = 1, 2, \ldots, 2^p$, where $\delta_{\mathbf{i}}$ are the p-tuples of the form $(\pm 1, \pm 1, \ldots, \pm 1)$. For a given β, these 2^p constrains can be split into two sets, one is a set of indices with equality constraints, namely, $E = \{i : \delta_{\mathbf{i}}^{\mathbf{T}}\beta = t\}$; the other one is a set of indices with inequality constraints, that is, $S = \{i : \delta_{\mathbf{i}}^{\mathbf{T}}\beta < t\}$. Start with $E = \{i_0\}$ where δ_{i0}=sign($\hat{\beta}^0$), $\hat{\beta}^0$ being the overall least squares estimate. It solves the least squares problem subject to $\delta_{\mathbf{i}}^{\mathbf{T}}\beta \leq t$, and then checks whether $\sum_{j=1}^{p} |\beta_j| \leq t$. If so, the computation is complete; if not the violated constraints is added to E and the process is continued until all 2^p constraints are satisfied. The procedure must always converge in the finite number of steps since one element is added to the set E at each step and there are a total of 2^p elements. By the Karush-Kuhn-Tucker (KKT) conditions, the final iterate is a solution to the original problem.

In the second algorithm which was suggested by Tibshirani (1996), each β_j is rewritten as $\beta_j^+ - \beta_j^-$, where β_j^+ and β_j^- are non-negative, and the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$ is equivalent to $\beta_j^+ \geq 0, \beta_j^- \geq 0$ and $\sum \beta_j^+ + \sum \beta_j^- \leq t$. In this way, the original problem with $p + 1$ variables and 2^p constraints is transformed to a new problem with more variables

$(2p + 1)$ but fewer constraints $(2p + 1)$. This new problem has the same solution as the original problem.

The above computation is performed for a fixed tuning parameter t. In Tibshirani (1996), fivefold cross-validation was suggested to estimate t. Suppose a regression model:

$$Y = \eta(X) + \epsilon,$$

where $E(\epsilon)$=0 and $var(\epsilon) = \sigma^2$. For an estimate $\hat{\eta}(X)$, the prediction error of $\hat{\eta}(X)$ is given by

$$\text{PE} = E\{Y - \hat{\eta}(X)\}^2.$$

The LASSO tuning parameter is standardized to $s = \frac{t}{\sum|\hat{\beta}_j^0|}$, where $\hat{\beta}_j^0 (j = 1, \cdots, p)$, is the ordinary least squares estimate. Then we estimate the prediction error over a grid values of s from 0 to 1 inclusively. The value of s which yields the lowest estimated prediction error PE is selected.

LASSO is a very efficient and popular method of variable selection, there are some other researchers studied the penalized likelihood methodology for variable selection in recent years (e.g., Fan and Liu (1999), Fan and Li (2001), Zou and Hastie (2005), Park and Hastie (2006), Fu (2007), Huang et. al. (2007)). Michael et. al. (2000) studied the properties of LASSO as a convex programming problem and derive its dual, an efficient algorithm for computing LASSO estimates was developed.

2.2.2 Some extension of LASSO

The LASSO is a penalized least square method imposing an L_1-penalty on the regression coefficient. However, as pointed by Zou and Hastie (2005), the LASSO has some limitations. First, in the case when $p > n$, the properties of the convex optimization problem make the LASSO select at most n variables. Also, the LASSO is not well defined unless the L_1 norm penalty is smaller than a certain value which is the ordinary least-squares estimate. Second, the LASSO can not deal with the high correlated grouped variables very well. This means it tends to select only one variable from the group and does not care which one is selected. To fix the above problems, Zou and Hastie (2005) proposed the Elastic Net, which adopts a combined L_1 and L_2 penalty, defined as follows:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{2.3}$$

This problem is equivalent to the optimization problem

$$\begin{aligned} \min_{\beta} \quad & \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{subject to} \quad & (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t. \end{aligned} \tag{2.4}$$

The function on the left hand side of the inequality constraint, namely,

$$(1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2$$

is called the *elastic net penalty*, which is a convex combination of the LASSO and ridge penalty. When $\alpha = 1$, Elastic Net then becomes the simple ridge regression and the

penalty becomes LASSO penalty if $\alpha = 0$.

Zou and Hastie (2005) proposed an efficient algorithm, namely LARS-EN to solve the above problem. LARS-EN is based on least angle regression (LARS) algorithm which is introduced in section 2.3. It can be proved that the elastic net problem is equivalent to a lasso problem on some augmented data, therefore LARS can be applied to solve the elastic net problem. In addition, since there are two tuning parameters in the elastic net, a cross-validation on a two-dimensional surface is used for choice of tuning parameter

Apart from the L_γ norm penalty, recently Fan and Li (2001) developed a penalized least square and likelihood approach with a modified penalty as follows:

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + n \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.5}$$

where p_λ is the penalty function which satisfies

$$p'_\lambda(|\beta_j|) = \lambda \left\{ I(|\beta_j| \leq \lambda) + \frac{(\alpha\lambda - |\beta_j|)_+}{(\alpha - 1)\lambda} I(|\beta_j| > \lambda) \right\} \text{ for some } \alpha > 2. \tag{2.6}$$

To solve the above problem, Fan and Li (2001) suggested a smoothing method to approximate the problem by a quadratic minimization problem and Newton-Raphson algorithm can be used. They also use the method of cross-validation to select the tuning parameter (λ, a).

2.2.3 Group LASSO

Consider the regression problem with J factors:

$$\mathbf{y} = \beta_0 + \mathbf{X}_1\beta_1 + \cdots + \mathbf{X}_J\beta_J + \epsilon,$$

where $\mathbf{X}_j$ is the factor with p_j grouped variables, $\beta_j = (\beta_{j1}, \cdots, \beta_{jp_j})'$ is a coefficient vector of size p_j, $j = 1, 2, \ldots, J$, and $\epsilon \sim N_n(0, \sigma^2 I)$. The J factors can be both categorical and continuous factors. Without loss of generality, the intercept β_0 can be set to be zero. It is easy to see that the general regression problem is the special case of the above when $p_1 = \ldots = p_J = 1$. Suppose we have data $(\mathbf{X_{ij}}, y_i), i = 1, 2, \ldots, n$, where $\mathbf{X_{ij}} = (X_{i1}, \ldots, X_{iJ})$. As proposed by Yuan and Lin (2006), the group LASSO estimate $(\hat{\beta}_1, \ldots, \hat{\beta}_J)$ is defined as the solution to the following constrained optimization problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{J}\beta_j X_{ij}\right)^2 \\ \text{s. t.} \quad & \sum_{j=1}^{J}\|\beta_j\|_{K_j} \le t, \end{aligned} \tag{2.7}$$

where the norm $\|\cdot\|_{K_j}$ is defined by $\|\beta_j\|_{K_j} := \sqrt{\beta_j' K_j \beta_j}$ for a symmetric positive definite matrix $K_j \in R^{p_j \times p_j}$. There are many reasonable choices for the kernel matrices K_js. In particular, an obvious and simple choice of K_j would be $K_j = I_{p_j}$, $j = 1, \ldots, J$, where I_{p_j} represents the identity matrix in $R^{p_j \times p_j}$.

Yuan and Lin (2006) proposed an optimization algorithm to obtain the solution of (2.7) under the selection of $K_j = p_j I_{p_j}, j = 1, \ldots, J$. By the direct consequence of

Karush-Kuhn-Tucker conditions, $\beta = (\beta_1', \ldots, \beta_J')'$ is a solution of (2.7) if and only if

$$-X_j'(Y - X\beta) + \frac{\lambda \beta_j \sqrt{p_j}}{\|\beta_j\|} \quad \text{for} \quad \beta_j \neq \mathbf{0}, \tag{2.8}$$

$$\| - X_j'(Y - X\beta)\| \leq \lambda \sqrt{p_j} \quad \text{for} \quad \beta_j = \mathbf{0}. \tag{2.9}$$

Hence if we further assume X is orthogonal, the solution can be obtained by solving the above expressions

$$\beta_j = \left(1 - \frac{\lambda \sqrt{p_j}}{\|S_j\|}\right)_+ S_j, \tag{2.10}$$

where $S_j = X_j'(Y - X\beta_{-j})$, with $\beta_{-j} = (\beta_1', \ldots, \beta_{j-1}', \mathbf{0}', \beta_{j+1}' \ldots, \beta_J')$.

2.3 Least angle regression (LARS) algorithm

To obtain the estimates of coefficients β in the regression (2.1), LARS performs analogously to a less greedy version of the forward selection approach. Starting with all coefficients equal to zero, the covariate which is most correlated with the response is selected next, say x_{j1}. Along x_{j1} the largest step is taken until some other covariate, say x_{j2}, has the same correlation with the current residual. Unlike the forward selection method, rather than continuing along x_{j1}, LARS takes the direction equiangular between the two covariates until the third variable x_{j3} attains the most correlation with the response. Then LARS proceeds equiangularly between x_{j1}, x_{j2} and x_{j3}, until the fourth variable is selected, and so on.

For a sample of size n, the response $\mathbf{y} = (y_1, \cdots, y_n)'$ and covariates $X_{n\times p} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$ can be always assumed that $\mathbf{y}$ has mean 0 and covariates $X_{n\times p}$ have mean 0 and unit

length by location and scale transformations. So, the intercept β_0 is zero. Let $\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}$, $i = 1, \cdots, n$ and denote $(\mu_1, \cdots, \mu_n)$ by $\boldsymbol{\mu}$. The LARS algorithm is described as follows:

I. Set the initial estimate $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and $k = 0$.

II. Suppose that $\hat{\boldsymbol{\mu}}_k$ is the current LARS estimate, calculate

$$\hat{\mathbf{c}} = X'(\mathbf{y} - \hat{\boldsymbol{\mu}}_k), \tag{2.11}$$

and set

$$\hat{C} = \max_j\{|\hat{c}_j|\}, \quad \text{and} \quad \mathcal{A} = \{j : |\hat{c}_j| = \hat{C}\}. \tag{2.12}$$

III. Let $X_{\mathcal{A}} = (\cdots \mathbf{x}_j \cdots)_{j \in \mathcal{A}}$ and calculate $\bar{\mathbf{y}}_{k+1} = (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} X'_{\mathcal{A}} \mathbf{y}$, then the equiangular vector

$$\boldsymbol{\mu}_{\mathcal{A}} = \bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k. \tag{2.13}$$

IV. The next step of LARS algorithm updates $\hat{\boldsymbol{\mu}}_k$ to

$$\hat{\boldsymbol{\mu}_{k+1}} = \hat{\boldsymbol{\mu}}_k + \hat{\gamma}\boldsymbol{\mu}_{\mathcal{A}}, \tag{2.14}$$

where

$$\hat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{\hat{C} - a_j}, \frac{\hat{C} + \hat{c}_j}{\hat{C} + a_j} \right\}, \tag{2.15}$$

and a_j is the jth element of $\mathbf{a} = X'_{\mathcal{A}} \boldsymbol{\mu}_{\mathcal{A}}$.

V. The process II-IV are repeated until $\mathcal{A}^c = \emptyset$.

The step length $\hat{\gamma}$ in (2.15) is the smallest positive number such that some new index in $\mathcal{A}^c$ joins the active set in next step.

2.4 Sparse logistic regression

Suppose the binary data $(\mathbf{X}_i, y_i), i = 1, 2, \ldots, n$, are n independent and identically distributed random vectors, where $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is the p-dimensional predictor variables and $y_i = 1$, or -1, is the binary response of interest. With the logit link function

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j,$$

where π_i is the probability of the response variable Y_i taking value 1, the negative log-likelihood is

$$-\sum_{i=1}^{n} \log\left(1 + \exp\left\{-y_i(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)\right\}\right). \tag{2.16}$$

By the idea of LASSO, we can solve the sparse logistic regression problem by solving the following optimization problem:

$$\min_{\beta} \quad l(\beta) = \sum_{i=1}^{n} \log\left(1 + \exp\left\{-y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})\right\}\right) + \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda > 0. \tag{2.17}$$

Shevade and Keerthi (2003) proposed an efficient algorithm to solve (2.17) which does not need any matrix operations but only the first order optimality conditions for (2.17).

Let

$$\begin{aligned} \eta_i &= -y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}), \\ H_0 &= \sum_{i=1}^{n} \frac{e^{\eta_i}}{1+e^{\eta_i}} y_i, \\ H_j &= \sum_{i=1}^{n} \frac{e^{\eta_i}}{1+e^{\eta_i}} y_i x_{ij}, \quad j = 1, 2, \ldots, p, \end{aligned}$$

since $l(\beta)$ is differentiable with respect to β_0, and β_j, $j > 0$ at the value of nonzero, the first order optimality conditions of (2.17) can be written as follows:

$$H_0 = 0, \text{ if } j = 0;$$

$$H_j - \lambda = 0, \text{ if } j > 0 \text{ and } \beta_j \neq 0;$$

$$|H_j| \leq \lambda, \text{ if } j > 0 \text{ and } \beta_j = 0.$$

Define viol_j, $j = 1, 0, 1, \ldots, p$, by

$$\text{viol}_j := \begin{cases} |H_0| & \text{if} j = 0 \\ |H_j - \lambda| & \text{if} j > 0 \text{ with } \beta_j \neq 0 \\ \max\{0, |H_j| - \lambda\} & \text{if} j > 0 \text{ with } \beta_j = 0. \end{cases} \tag{2.18}$$

Then the first order optimality conditions of (2.17) can be further written as

$$\text{viol}_j = 0, \quad \text{for any } j.$$

In Shevade and Keerthi (2003), a tolerance τ is suggested in a algorithm used to solve (2.17) such that the the algorithm stops when the optimality conditions are satisfied up to tolerance τ, i.e.

$$\text{viol}_j \leq \tau, \quad \text{for any } j.$$

Under a given tuning parameter λ, Shevade and Keerthi (2003) proposed an algorithm to solve (2.17) as follows, all βs are set to zero initially, one variable β_j which maximumly violates the optimality conditions is chosen and the optimization subproblem is solved with respect to this variable β_j alone, keeping the other βs fixed. This procedure is repeated as long as there exists a variable which violates the optimality conditions.

Chapter 3

TDT for Quantitative Traits

Original TDT concentrated on dichotomous traits such as the presence or absence of a disease. Generally, the disease was caused by the mutation at a single gene. But as we have known, many biomedical traits of interest can be measured on a continuous or ordinal scale, namely quantitative scale. A quantitative trait is usually attributable to two or more genes as well their interaction with the environment. So, it is important to study linkage between a marker locus and the locus underlying such a quantitative trait. In this chapter we will first review some available methods which extend the application of TDT to quantitative trait locus (QTL) mapping, then we provide a new sampling method, namely extreme rank sampling (ERS), which can convert the quantitative trait study to the case-control like study. Some properties of ERS are studied. The TDT is extended to QTL mapping by applying ERS. Simulation studies are carried out to compare the powers of TDT with ERS with conventional truncation sampling scheme.

3.1 Existing methods of TDT for QTL mapping

Since Spielman et. al. (1993) developed TDT for mapping disease susceptibility gene at a single locus, some other researchers have proposed various TDT-type procedure for QTL mapping. We review these methods in this section.

Throughout this chapter, we assume that the population is in Hardy-Weinberg equilibrium. Let X be the quantitative trait of concern which is controlled by a putative biallelic QTL A with alleles Q and q. Assume Q is related to larger trait value. Let p_Q denote the frequency of the Q-allele, $p_q(= 1 - p_Q)$ denote the frequency of the q-allele. Suppose that a genetic marker locus B is in the vicinity of the QTL A and let p_M and $p_m(= 1 - p_M)$ denote the frequencies of the allele M and m at the marker locus, respectively. Suppose that the marker B is in linkage disequilibrium (LD) with QTL A, and allele M is linked with Q allele– that means M is associated with larger trait values.

3.1.1 t-test with random sampling

The t-test TDT approach developed by Allison (1997) is quite different from the TDT as originally proposed by Spielman et. al. (1993). Assuming one child per family, this method apply to the family trios consisting one heterozygous parent and one homozygous parent, and their biological child. Under this condition, we would be very clear which of the two alleles at the marker locus is transmitted from the heterozygous parent to the child. Let $\bar{X}_M$ and $\bar{X}_m$ denote the sample means of quantitative trait value

of children, and S_M^2 and S_m^2 denote the sample standard deviations when allele M or m is transmitted, respectively. Given the sample size N, the test statistic

$$\text{TDT}_{t1} = \frac{\bar{X}_M - \bar{X}_m}{\sqrt{\frac{2(S_M^2 + S_m^2)}{N}}}$$

will have t-distribution with $N - 2$ degree freedom under the null hypothesis of no linkage. The t-statistic above assume either that the quantitative trait X is normal or that the sample size is large such that the central limit theorem applies.

3.1.2 t-test with truncation sampling

This is a TDT based on t-test with truncation sampling from both tails of the offspring phenotype distribution. For the given lower and upper cut off points X_L and X_U, where $X_L \leq X_U$, the families in which the quantitative value X of child is less than X_L or greater than X_U are selected and are put into lower and upper groups. Then,under the assumption that the sample size is sufficiently large such that the central limit theory can be applied. The statistic

$$\text{TDT}_{t2} = \frac{\bar{X}_M - \bar{X}_m}{\sqrt{\frac{S_M^2}{N_M} + \frac{S_m^2}{N_m}}}$$

is t-distribution with degree freedom of $N_M + N_m - 2$, where $\bar{X}_M$, $\bar{X}_m$, S_M^2 and S_m^2 are the same as that in TDT_{t1}, N_M and N_m are the number of families where allele M or m is transmitted, respectively.

3.1.3 F-test for linkage

Without assuming only one parent to be heterozygous only one child in a family, Xiong et. al. (1998) proposed a general TDT for quantitative traits. For a random sample consisting of N families which there is at least one parent who is heterozygous for a marker locus, there are $2N$ sampled parents. For the k-th heterozygous parent, assume allele M is transmitted to C_{Mk} children and allele m is transmitted to C_{mk} children. Let $N_M = \sum_k C_{Mk}$ and $N_m = \sum_k C_{mk}$. For the j-th child in the set of C_{Mk} children and j-th child in the set of C_{mk} children, let X_{Mkj} and X_{mkj} denote the quantitative trait value respectively. The TDT statistic for testing the null hypothesis of no linkage is

$$\mathrm{TDT}_F = \frac{(\bar{X}_M - \bar{X}_m)^2}{\left(\frac{1}{n_M} + \frac{1}{n_m}\right) S^2},$$

where

$$\bar{X}_M = \frac{1}{N_M} \sum_k \sum_j X_{Mkj}, \qquad \bar{X}_m = \frac{1}{N_m} \sum_k \sum_j X_{mkj},$$

and

$$S^2 = \frac{\sum_k \sum_j [(X_{Mkj} - \bar{X}_M)^2 + (X_{mkj} - \bar{X}_m)^2]}{N_M + N_m - 2}.$$

Under the assumption that the quantitative trait is normally distributed and the null hypothesis that there is no linkage, the TDT_F follows an F distribution with 1 and $N_M + N_{m-2}$ degrees of freedom. When the sample size is sufficiently large, even without normality assumption the TDT_F will have $\chi^2_{(1)}$ distribution asymptotically.

3.1.4 χ^2-test with truncation sampling

Assume the phenotype has normal distribution, a lower point X_L and an upper point X_U for phenotype of children are given first. As proposed by Allison (1997), the families in which the quantitative value X of child is less than X_L or greater than X_U are selected and are put into lower and upper groups respectively. Then the transmission status of M allele and m allele can be summarized in the Table (3.1).

Table 3.1: Number of alleles M and m transmitted in upper sample and lower sample

	Number of alleles transmitted		
	M	m	Total
Upper sample	N_M^U	N_m^U	n^U
Lower sample	N_M^L	N_m^L	n^L
Total	n_M	n_m	N

To test whether the transmission of putative allele M is independent of higher phenotype value, a statistic is as follows:

$$\text{TDT}_{chi} = \frac{N(N_M^U N_m^L - N_M^L N_m^U)^2}{n_M n_m n^U n^L}.$$

Under the null hypothesis of no linkage, the above statistic has χ^2 distribution with degree freedom 1. As pointed by Allison (1997), such test is advantageous if segregation distortion is considered because both higher and lower phenotypes are considered in the statistic.

Among these 4 tests reviewed, it can be seen that only the last one retains the spirit of TDT whereas the first three only use the family trios and the transmission data, but has nothing to do with the spirit of TDT.

3.2 A new sampling method and its properties: extreme rank sampling for TDT in QTL mapping

In this section we apply an extreme rank selection (ERS) (Chen et. al. (2005)) procedure to selective genotyping such that TDT can be used in QTL mapping. Some properties of ERS in TDT are studied. By simulation studies, we show that, while having the comparable power with the truncation approach, ERS approach can improve efficiency of sampling and therefore reduce cost .

ERS approach is applied only to phenotypes of children and each of them is from different nuclear family. Let k be a specified integer. By ERS approach, at each step, k children from k different families are chosen at random. The trait values of these k children are measured and ordered from the smallest to the largest: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(k)}$. Then the parents of the child with rank 1 and of the child with rank k are genotyped at the marker locus. If there is at least one heterozygous parent for the family of rank 1 child, then this family is selected as a member of the lower sample. Similarly, If there is at least one heterozygous parent for the family of rank k child, then this family is selected as a member of the upper sample. That is, the lower sample contains lower

child-parent trios while the upper sample contains upper child-parent trios.

Denote the three genotypes of A by QQ Qq and qq. Let p_{QQ}, p_{Qq}, p_{qq} denote the frequencies of the genotype QQ, Qq, and qq, respectively. Under Hardy-Weinburg-Equilibrium (HWE),

$$p_{QQ} = p_Q^2, \quad p_{Qq} = 2p_Q p_q, \quad \text{and} \quad p_{qq} = p_q^2.$$

By quantitative genetics, without loss of generality it is postulated that

$$X = g + \epsilon,$$

where g is the genotypic value and ϵ is a random variable independent of X with mean μ and cumulative distribution function (CDF) H.

Assume that the genotypic value g is a, d and $-a$ when the genotypes of the QTL are QQ Qq and qq respectively. Then

$$F_{qq}(x) = H(x + a), \quad F_{Qq} = H(x - d), \quad \text{and} \quad F_{QQ} = H(x - a)$$

are CDF of X for different genotypes of QTL. Denote by f_{qq}, f_{Qq} and f_{QQ} their probability density functions (PDF). Thus, the PDF and CDF of X is given, respectively, by

$$\begin{aligned} f(x) &= p_Q^2 f_{QQ}(x) + 2p_Q p_q f_{Qq}(x) + p_q^2 f_{qq}(x), \\ F(x) &= p_Q^2 F_{QQ}(x) + 2p_Q p_q F_{Qq}(x) + p_q^2 F_{qq}(x), \end{aligned}$$

and the conditional frequencies of the genotype QQ, Qq and qq given X are as follows:

$$p(QQ|X = x) = \frac{p_Q^2 f_{QQ}(x)}{f(x)}, \quad p(Qq|X = x) = \frac{2p_Q p_q f_{Qq}(x)}{f(x)}, \quad p(qq|X = x) = \frac{p_q^2 f_{qq}(x)}{f(x)}.$$

Let δ denote LD coefficient for the marker locus and the QTL, and θ denote the recombination fraction between the marker locus and QTL. Then, the haplotype frequencies are as follows.

$$p_{QM} = \delta + p_Q p_M, \quad p_{qM} = p_M - p_{QM}, \quad p_{Qm} = p_Q - p_{QM}, \quad \text{and} \quad p_{qm} = 1 - p_Q - p_M + p_{QM}.$$

In the following we will prove that the probability that a selected parent is heterozygous at the marker locus (i.e., genotype is Mm) and the M allele is transmitted to his/her child can be increased in upper sample and reduced in lower sample by ERS approach.

Let the subscripts C and P denote the child and parental generations, two letters, q and m, jointing with a underline represent a haplotype of marker locus and QTL, and P_M and P_m the allele frequencies of M and m, respectively. So the population frequency of the genotype Mm_P in a parent and the allele M transmitted to the child is

$$P(Mm_P, M_C) = P(M_C|Mm_P)P(Mm_P) = P_M P_m.$$

Considering three cases of genotype of QTL, we have

$$\begin{aligned}
P(Mm_P, \underline{qM}_C) &= P[(\underline{QM}, \underline{qm})_P, \underline{qM}_C] \\
&\quad + P[(\underline{qM}, \underline{Qm})_P, \underline{qM}_C] \\
&\quad + P[(\underline{qM}, \underline{qm})_P, \underline{qM}_C] \\
&= 2P_{QM}(1 - P_M - P_Q + P_{QM})\frac{\theta}{2} \\
&\quad + 2(P_M - P_{QM})(P_Q - P_{QM})\frac{1-\theta}{2} \\
&\quad + 2(P_M - P_{QM})(1 - P_M - P_Q + P_{QM})\frac{1}{2} \\
&= \delta(\theta - P_m) + P_M P_m P_q.
\end{aligned}$$

Similarly,

$$P(Mm_P, \underline{QM}_C) = \delta(P_m - \theta) + P_M P_m P_Q.$$

Therefore, the joint frequency of Mm_P and the allele M transmitted to a child given the quantitative trait value X of the child is:

$$\begin{aligned}
P(Mm_P, M_C|X) &= P(Mm_P, \underline{qM}_C, X = x)/f(x) + P(Mm_P, \underline{QM}_C, X = x)/f(x) \\
&= \{P(Mm_P, \underline{qM}_C, Qq_C)f_{Qq}(x) + P(Mm_P, \underline{qM}_C, qq_C)f_{qq}(x) \\
&\quad + P(Mm_P, \underline{QM}_C, QQ_C)f_{QQ}(x) + P(Mm_P, \underline{QM}_C, Qq_C)f_{Qq}(x)\}/f(x) \\
&= \{P_Q P(Mm_P, \underline{qM}_C)f_{Qq}(x) + P_q P(Mm_P, \underline{qM}_C)f_{qq}(x) \\
&\quad + P_Q P(Mm_P, \underline{QM}_C)f_{QQ}(x) + P_q P(Mm_P, \underline{QM}_C)f_{Qq}(x)\}/f(x) \\
&= [\delta(\theta - P_m) + P_M P_m P_q][P_Q f_{Qq}(x) + P_q f_{qq}(x)]/f(x) \\
&\quad + [\delta(P_m - \theta) + P_M P_m P_Q][P_Q f_{QQ}(x) + P_q f_{Qq}(x)]/f(x).
\end{aligned}$$

Thus, the joint probability that the parent has genotype Mm_P and the child has allele M_C in upper sample $P^U(Mm_P, M_c)$ can be expressed as:

$$\begin{aligned}
& E[P(Mm_P, M_c|X_{(k)})] \\
= & \int kF^{k-1}(x)\{[\delta(\theta - P_m) + P_M P_m P_q][P_Q f_{Qq}(x) + P_q f_{qq}(x)] \\
& + [\delta(P_m - \theta) + P_M P_m P_Q][P_Q f_{QQ}(x) + P_q f_{Qq}(x)]\}dx \\
= & P_M P_m \int kF^{k-1}(x)[P_Q^2 f_{QQ}(x) + 2P_Q P_q f_{Qq}(x) + P_q^2 f_{qq}(x)]dx \\
& + \delta(P_m - \theta) \int kF^{k-1}(x)[P_Q(f_{QQ}(x) - f_{Qq}(x)) + P_q(f_{Qq}(x) - f_{qq}(x))]dx \\
= & P_M P_m + \delta(P_m - \theta) \int kF^{k-1}(x)[P_Q(f_{QQ}(x) - f_{Qq}(x)) + P_q(f_{Qq}(x) - f_{qq}(x))]dx.
\end{aligned} \tag{3.1}$$

Similarly, replacing $X_{(k)}$ and $F(x)$ by $X_{(1)}$ and $1 - F(x)$, we obtain

$$\begin{aligned}
& E[P(Mm_P, M_c|X_{(1)})] \\
= & \int k[1-F(x)]^{k-1}\{[\delta(\theta - P_m) + P_M P_m P_q][P_Q f_{Qq}(x) + P_q f_{qq}(x)] \\
& +[\delta(P_m - \theta) + P_M P_m P_Q][P_Q f_{QQ}(x) + P_q f_{Qq}(x)]\}dx \\
= & P_M P_m \int k[1-F(x)]^{k-1}[P_Q^2 f_{QQ}(x) + 2P_Q P_q f_{Qq}(x) + P_q^2 f_{qq}(x)]dx \\
& +\delta(P_m - \theta)\int k[1-F(x)]^{k-1}[P_Q(f_{QQ}(x) - f_{Qq}(x)) + P_q(f_{Qq}(x) - f_{qq}(x))]dx \\
= & P_M P_m + \delta(P_m - \theta)\int k[1-F(x)]^{k-1}[P_Q(f_{QQ}(x) - f_{Qq}(x)) + P_q(f_{Qq}(x) - f_{qq}(x))]dx.
\end{aligned} \tag{3.2}$$

Slatkin (1999) suggest that we can usually assume that frequency of allele M is sufficiently small, so $P_m > 1/2$. Because the recombination coefficient $\theta \leq 1/2$, while the integral in last part of (3.1) is positive. Thus, it follows from (3.1) that

$$P^U(Mm_P, M_c) \geq P(Mm_P, M_c),$$

which implies that the ERS can increase the joint frequency of trio that at least one of parents is heterozygous and a child has a M allele. $[1 - F(x)]^{k-1}$ is decreasing so the integral in last part of (3.2) is negative. This leads to a reduced joint frequency $P(Mm_P, M_c)$ in lower sample. In other words, for the same sample size, compared with random sampling, we can expect more family trios heterozygous parents at the marker locus and allele M is transmitted to the child in upper sample and meanwhile less of family trios with heterozygous parents at the marker locus and allele M is transmitted to the child in lower sample. Therefore, by ERS approach we can improve the efficiency

of sampling and consequently improve the power of TDT test based on upper and lower samples. In the following section, we introduce some statistical tests employing the upper and lower samples obtained by ERS to identify QTL.

3.3 TDT for QTL mapping by ERS

Analogous to Allison's χ^2 test with truncation sampling approach, we will consider three test statistics in our discussion. However, we don't impose the restriction that family trios must consist of one heterozygous and one homozygous parent. Instead, we consider the family trios with at least one heterozygous parent. Let N_M^U and N_m^U denote, respectively, the number of times that allele M and m at marker locus are transmitted from heterozygous parents to the children with rank k in the upper sample obtained by ERS, the detailed notations are shown in Table 3.1. Here the lower and upper sample are selected by ERS approach instead of truncation sampling method. Then under null hypothesis that there is no linkage and no linkage disequilibrium between marker locus and a QTL, the statistic

$$\chi^2_{TDT_U} = \frac{(N_M^U - N_m^U)^2}{N_M^U + N_m^U}$$

approximately follows a χ^2 distribution with one degree of freedom.

Similarly, for lower sample, let N_M^L and N_m^L denote, respectively, the number of times that allele M and m at marker locus are transmitted from heterozygous parents to the children with rank 1 in the lower sample. Then under null hypothesis that there is no

linkage and no linkage disequilibrium between marker locus and a QTL, the statistic

$$\chi^2_{TDT_L} = \frac{(N^L_M - N^L_m)^2}{N^L_M + N^L_m}$$

approximately follows a χ^2 distribution with one degree of freedom as well.

Spielman et al. (1993) suggested that we should incorporate unaffected offsprings into the original TDT if we have concern about the effect of segregation distortion rather than linkage disequilibrium on *prob*(M is transmitted). Thus we have the following statistic which combines the upper and lower samples together,

$$\chi^2_{TDT_{UL}} = \frac{N(N^U_M N^L_m - N^L_M N^U_m)^2}{n_M n_m n^U n^L}.$$

This statistic approximately follows a χ^2 distribution with one degree of freedom under the null hypothesis and it may be advantageous when the possibility of segregation distortion exits. These three tests are referred to as TDT_u, TDT_l and TDT_{ul}, respectively.

The above three TDT statistics consider only biallelic genetic model. For the locus with more than two alleles, as pointed out by Deng and Chen (2001), it can always be classified into two alleles by designating one or some alleles as M and the others as m. But in practice, collapsing multiple alleles to two compound alleles is not always straightforward since which alleles are to be grouped into one allele is not explicit. Inappropriate collapsing may lead to loss of information. Some researchers studied the extension of TDT for biallelic locus to multiallelic locus (e.g Sham And Curtis 1995, Spielman and Ewens 1996). We can use multiple comparison procedure to test the overall linkage disequilibrium among all of the alleles.

3.4 The power of TDT with ERS

In this section, we perform simulations and investigate powers of the various TDTs with ERS approach introduced in the above section and compare them with TDT under the truncation sampling scheme. With specified allele frequencies of QTL and maker locus p_Q, p_M,and linkage disequilibrium coefficient δ which is determined by $D' = \delta/[p_Q(1-p_M)]$, the parents' genotypes are generated randomly by HWE; in the absence of segregation distortion, with a specified recombination fraction θ, the genotype of their unique child is generated by the Mendelian inheritance. Furthermore, we assume that the effects of QTL alleles are additive, i.e., the distribution of trait X has mean 0, d, and 2d when the genotype of QTL is qq, Qq, and QQ, respectively. To make the results comparable, we conduct TDT_U, TDT_L and TDT_{UL} with the same sample size $n = 200$. In TS approach, let α_u and α_l to be the τ-th quantile and $(1-\tau)$-th quantile of the distribution of trait value of children. Correspondingly, in ERS approach we take k to be $\frac{1}{\tau}$. However, in practice, we don't know the the distribution of trait value, so we need to sample from the population of the trait value to estimate α_u and α_l first. To make the implementation easier and reduce the cost in practice, we also consider the idea of weaker version of ERS; that is, in the case that none of the parents of the rank k child is heterozygous at the marker locus we alternately choose the trio of the rank $k-1$ child in ERS sampling, and similarly for the family of the rank 1 child. Thereafter, the original ERS and the one under the weaker version will be referred to as ERSI and ERSII.

In the simulation study, the samples of size n for TS and ERS approaches are generated as follows. In TS approach, the population with the genotypes at the QTL and marker locus under the HWE is firstly generated. Then, *nk* trait values are randomly generated based on the variant distributions of the trait value and are used to estimate upper and lower quantile α_u and α_l. The genotypes of the children are generated according to the Law of Mendelian and denote the copies of family trios by (X, g_c, g_m, g_f), where X is the trait value of the child, g_c, g_m and g_f are the genotypes of child, mother and father respectively at the QTL and the marker locus. If X is larger than the upper quantile α_u, and at least one of the parents is heterozygous at marker locus, then this family is selected to be a member of upper sample for TDT_U. Similarly, If X is less than the lower quantile α_l, and at least one of the parents is heterozygous at marker locus, then this family is selected to be a member of lower sample for TDT_L. If we simultaneously consider whether X is greater than the upper quantile α_u or X is smaller than the lower quantile α_l and make the sum of the upper members and the lower members to be n, then a sample for TDT_{UL} is obtained.

For the ERS approach, a batch of k copies of families is generated and the trait values X's of children are ordered from smallest to largest: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(k)}$, then we observe the genotypes of parents of rank k and rank 1 children, if at least one of parents is heterozygous at the marker locus in a family, then a member of the upper sample or of the lower sample is obtained. Repeat this process, until we get a upper sample for TDT_U and a lower sample for TDT_L or a sample for TDT_{UL}.

The frequency of the Q allele and M allele is taken to be 0.02, 0.03 respectively. δ is the linkage disequilibrium coefficient, it is determined by normalized disequilibrium coefficient $D' = \delta/[P_Q(1-P_M)]$. To assess type I errors, we simulated data with $\theta = 0.5$, $D' = 0$ and $d = 0$. For each set of parameter values, we generate 1,000 replicates of samples by ERS and TS approaches to compare the powers of three tests— TDT_U, TDT_L and TDT_{UL} under truncation sampling approach (TS) and ERS. The proportion of rejections of each test with the same approach among the 1000 replicates is counted. In the case $d = 0$, $D' = 0$ and $\theta = 0.5$, this proportion provides an approximation of the probability of type I error. In the case $d \neq 0, D' \neq 0, \theta < 0.5$, this proportion provides the approximation of the power of the test. Some simulation results are reported in Table 3.2 and illustrated by Figure 3.1 and Figure 3.2.

It is noted that for specified d and D', the power of each TDT decrease with increasing θ, this is because the bigger the value of θ the weaker linkage of QTL and the marker locus. And meanwhile, with the same settings of parameters, under the same sampling approach, TDT_U is always more powerful than TDT_{UL} and TDT_L. This is demonstrated in Figure 3.2. If the effect of combination fraction θ and the genetic effect on quantitative trait remain constant, then with the linkage disequilibrium coefficient decreasing the power of each TDT decreases. In table 3.2, we also found the effect of heritability on the power of all TDT because the increasing heritability is due to the increase in d with the fixed allele frequency of QTL and fixed environment effect, the increases with the increasing heritability. Furthermore, for the same parameter set, from Figure 3.1 we found that TDT under truncation sampling approach is always slightly more powerful

Table 3.2: With the significant level $\alpha = 0.05$, power comparison of TDT under TS, ERSI and ERSII sampling approaches

Parameter Values			TDT_U			TDT_{UL}			TDT_L		
d	D'	θ	TS	ERSI	ERSII	TS	ERSI	ERSII	TS	ERSI	ERSII
0	0	0.5	0.046	0.059	0.053	0.062	0.058	0.055	0.055	0.052	0.055
0.8	0.85	0.01	1.00	0.997	0.972	1.00	0.989	0.949	0.988	0.976	0.958
0.8	0.85	0.05	0.998	0.996	0.934	0.978	0.971	0.901	0.977	0.949	0.901
0.8	0.85	0.09	0.995	0.986	0.900	0.966	0.939	0.859	0.950	0.916	0.841
0.8	0.45	0.05	0.877	0.785	0.545	0.700	0.597	0.448	0.448	0.383	0.358
0.8	0.70	0.05	0.991	0.970	0.859	0.933	0.898	0.792	0.849	0.804	0.723
0.8	0.95	0.05	1.00	0.999	0.969	0.998	0.985	0.967	0.994	0.989	0.971
0.4	0.85	0.05	0.714	0.630	0.489	0.695	0.595	0.451	0.624	0.528	0.446
0.61	0.85	0.05	0.969	0.924	0.794	0.944	0.944	0.767	0.897	0.822	0.747
1.03	0.85	0.05	1.00	1.00	0.993	1.00	1.00	0.982	0.989	0.981	0.981

than the ERS approach.

Although TS approach is slightly powerful than ERSI and ERSII, TS is more difficult to implement in practice, because a pre-screening process is necessary for the estimation of the cutoff quantile if they are not known a priori, which is usually the case. Generally, the pre-screening is a time-consuming and cost-added process. For example, Xu et.al (1999) pre-screened 40,000 individuals to estimate the cutoff quantiles of blood pressure in their study of mapping genes that regulate blood pressure. In contrast, ERS method does not need a pre-screening process. The selection is done in

batches of k individuals and the number k is usually small, this makes ERS more manageable. Thus, in the case when a large pre-screening process is needed, ERS approach is an alternative because of its convenience of implementation.

In the comparison of power of variant TDTs under the same sampling approaches, we find that under the condition that the allele Q has the positive effect on the quantitative trait, i.e., $d > 0$, TDT_u is always most powerful. In the case of disease susceptibility gene mapping, the samples corresponding to upper and and lower samples are case and control samples, Deng and Chen (2001) investigate the power of three TDT statistics analogous to TDT_u, TDT_l and TDT_{ul}, with larger genetic effect or larger prevalence of disease, or larger frequency of disease allele the TDT which incorporates unaffected offspring is more likely to be more powerful than original TDT. By our simulation study, we found that in QTL mapping by TDT with ERS, the genetic effect and frequency of allele related to the increasing QT value are compounded in heritability so we compare TDT_u, TDT_l and TDT_{ul} under various heritabilities of the quantitative trait. On the hand, we consider the fixed small frequency of QTL alleles, in this case TDT_u is more likely to be more powerful than the others, this is consistent with the conclusion of Deng and Chen (2001).

In the above simulations, the batch size k of ERS is chosen to be 10. It is apparently that with the increasing in k, the the tests are more powerful with the fixed sample size. However k can not to be chosen too big. Lander and Botstein (1989) warned that very extreme trait values might result from other causes rather than genetic effects. They

suggested that for the truncation approach the upper and the lower percentage should not be less than 5 percentage. That means the batch size k should not be more than 20.

Figure 3.1: Power comparison of ERSI, ERSII and TS sampling approaches

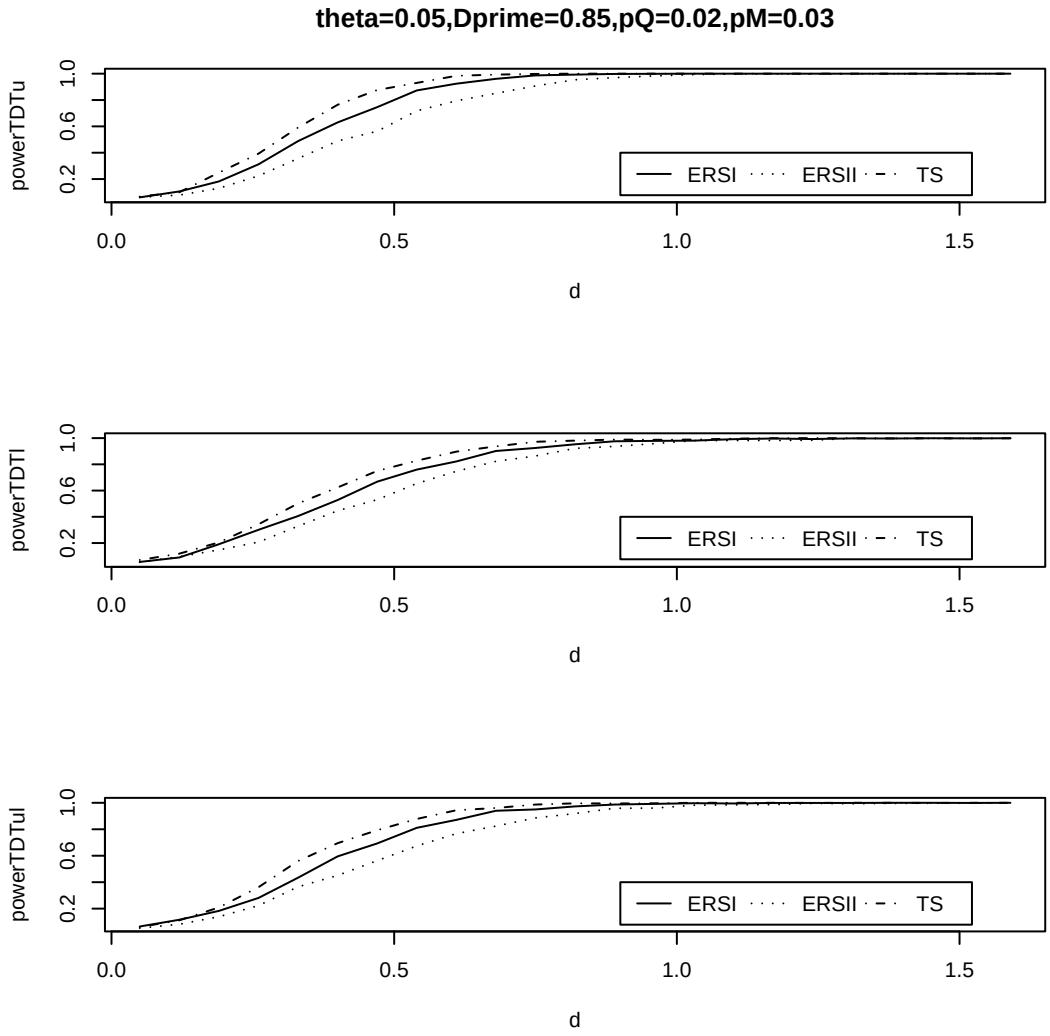
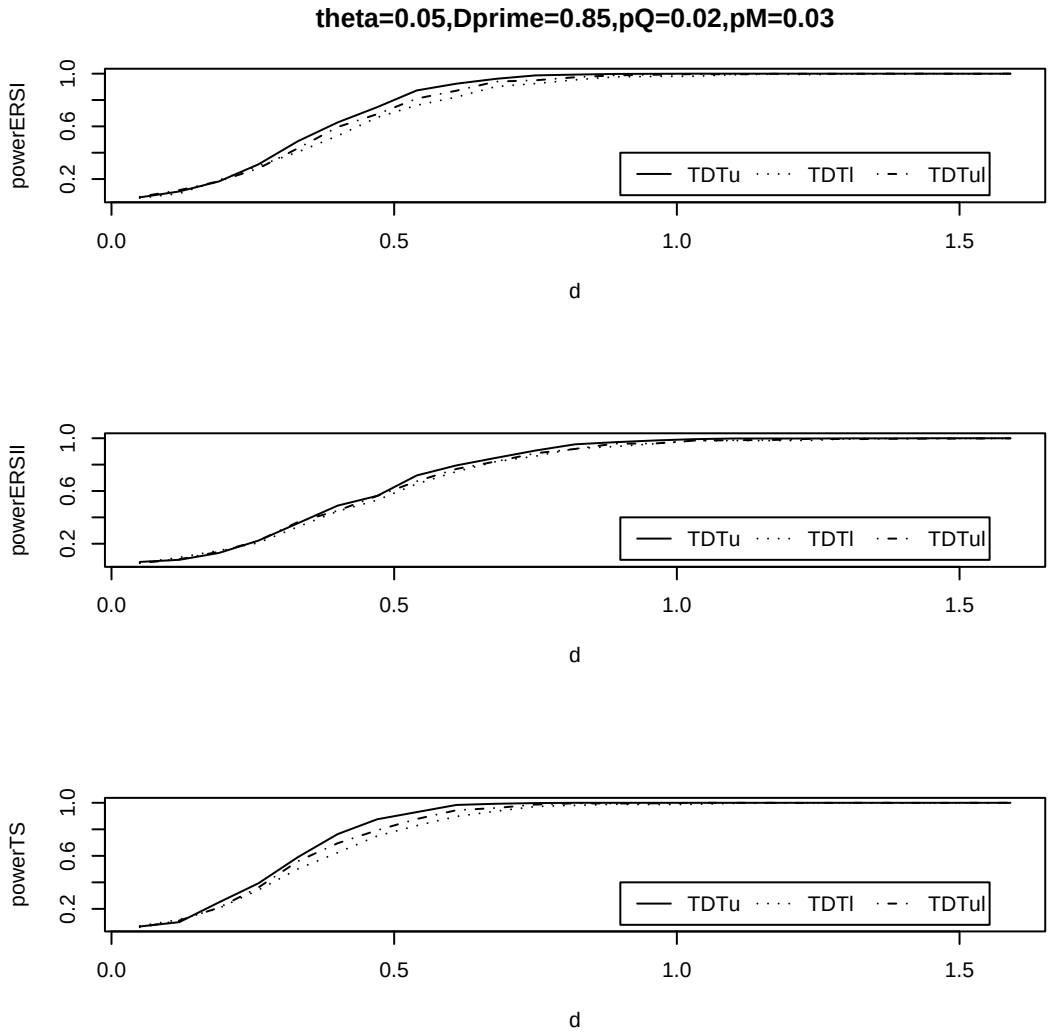


Figure 3.2: Power comparison of TDT_U, TDT_L and TDT_{UL} under the same sampling approach



Chapter 4

TDT in Genome-wide Association Study

With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, together with the rapid improvements in SNP genotyping technology, nowadays researchers have a set of research tools available that make it possible to find genetic variants to common diseases. We have been concerned with linkage analysis by TDT on a single locus. This single-locus TDT has been applied to genome-wide studies by using Bonferroni adjustment. In this chapter, we provide a new approach to extend the application of the TDT to genome-wide association studies by using logistic models and some selection criteria. By simulation studies, we compare our approach with the traditional Bonferroni-type procedures.

4.1 FDR-controlling procedure

When the TDT was introduced by Spielman et al (1993), the intended use was as a test for linkage with a particular marker which is very near a candidate gene in the case that the disease association has already been found. However, even if prior evidence for association is absent, the TDT is still valid to test the linkage between markers and disease susceptibility genes by linkage disequilibrium study. For the case of multilocus, Spielman et al (1993) suggested a standard Bonferroni correction for multiple independent tests, this procedure is used to control type I error. With the developing of SNP genotyping technology, people can locate a large amount of SNP markers throughout the genome. If the number of comparison is very large, the standard Bonferroni correction approach is very conservative because of the large critical value used. Benjamini et. al. (1995) developed a less conservative approach, namely, false discovery rate controlling procedure (FDRC). In this section, we introduce Bonferroni approach and FDRC method.

Suppose we conduct k (k is usually large) dependent or independent statistical tests which are being performed simultaneously on a set of data with significant level $\alpha = 0.05$, then there will be 1 out of every 20 hypothesis-tests appearing to be significant merely due to chance. In order to avoid spurious false positives, Bonferroni adjusted values are used in multiple tests. The Bonferroni adjustment sets the critical value for each individual test at level α/k in order to control the overall false positive rate at level α.

As Spielman et. al. (1996) pointed out, the Bonferroni correction procedure is quite conservative, especially when the number of markers is large (it is always the case for genome-wide studies), the significant level will be very extreme, so, it is likely that very few of the hypotheses would be rejected. This results in a great loss in power. Instead, an alternative approach based on false-discovery-rate (FDR) called *false discovery rate controlling procedure* was developed by Benjamini and Hochberg (1995). Instead of controlling the overall type I error, the FDR controlling procedure attempts to control the proportion of falsely rejected null hypotheses among all rejected null hypothesis. Suppose that we have conducted k tests $H_1, H_2, \ldots, H_k$ with the corresponding p-values $p_1, p_2, \ldots, p_k$. The p-values are ordered as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$ and the null hypotheses corresponding to $p_{(i)}$ is denoted by $H_{(i)}$. For a given α, let

$$s = \arg\max_i \left\{ p_{(i)} \leq \frac{i}{k}\alpha \right\},$$

then all $H_{(i)}$, $i = 1, 2, \ldots, s$, will be rejected. In Benjamini et. al. (1995), the above procedure was proved to be able to control the FDR at α.

Table 4.1: Alleles transmission in case group (size is N_1) and control group (size is N_2) at p SNP marker loci

	No. of alleles transmitted				
	1	$\cdots$	j	$\cdots$	p
Case	(n_M^1, n_m^1)	$\cdots$	(n_M^j, n_m^j)	$\cdots$	(n_M^p, n_m^p)
Control	(u_M^1, u_m^1)	$\cdots$	(u_M^j, u_m^j)	$\cdots$	(u_M^p, u_m^p)
Subtotal	(t_M^1, t_m^1)	$\cdots$	(t_M^j, t_m^j)	$\cdots$	(t_M^p, t_m^p)

In this chapter, suppose that we have genotypes of case-parent trios and control-parent trios at a large number of biallelic SNP markers spread throughout the genome where only one affected or unaffected child from each family is employed. At each marker locus, assume two alleles are M and m, the data of allele transmission is summarized in Table (4.1). At jth maker locus, under the null hypothesis of no linkage or no linkage disequilibrium, we consider three types of TDT statistics described in the following.

(1) TDT_1, it is applied to case-parent trios only, the statistic

$$\chi^2_{TDT_1} = \frac{(n^j_M - n^j_m)^2}{n^j_M + n^j_m} \tag{4.1}$$

approximately follows a χ^2 distribution with degree freedom 1.

(2) TDT_2, it is applied to control-parent trios only, the statistic is

$$\chi^2_{TDT_2} = \frac{(u^j_M - u^j_m)^2}{u^j_M + u^j_m}, \tag{4.2}$$

and approximately follows a χ^2 distribution with degree freedom 1.

(3) TDT_3, it is applied to both case-parent and control-parent trios, the statistic is defined as

$$\chi^2_{TDT_3} = \frac{(N_1 + N_2)(n^j_M u^j_m - n^j_m u^j_M)^2}{t^j_M t^j_m N_1 N_2}, \tag{4.3}$$

TDT_3 approximately follows a χ^2 distribution with degree freedom 1. The power of these TDTs were compared in detail by Deng et. al. (2001).

In genome-wide association study, we might have thousands or tens of thousands potential makers available. But the sample size is relatively small (only tens or hundreds of subjects) i.e., p is much larger than sample size n and only a small number of them associate with the disease, thus we should conduct a large number of tests so we use FDRC procedure to control false-positive results. With one of the above TDT statistics, we firstly calculate p-value for each of the markers loci, then under a critical level (usually 0.05) we apply FDRC procedure to test the null hypothesis that no linkage or linkage disequilibrium between a marker locus and the disease susceptibility locus, the markers at which we reject the null hypothesis are considered to be linked with the disease susceptibility locus. Thereafter, we use FDRC-TDT1, FDRC-TDT2 and FDRC-TDT3 to denote these methods respectively.

4.2 Genome-wide TDT procedure using logistic model and feature selection techniques

At each marker locus, by genetics, it is known that there are four situations of transmission of alleles from parents to the child: (*i*) MM parent transmits M to child, (*ii*) mm parent transmits m to child, (*iii*) Mm parent transmits M to child, and (*iv*) Mm parent transmits m to child. Based on the genotypes of the parents and child, we can obtain the information of transmission of each allele from parents to offsprings. In the absence of segregation distortion, in nuclear families with affected offspring and with

at least one parent heterozygous for a marker, individual alleles should be transmitted from heterozygous parents to affected child with equal probabilities unless the maker locus is linked to and in linkage disequilibrium with a disease susceptibility locus or is a disease susceptibility locus itself. Thus, we can model the probability of having the disease for a child by logistic regression model with the numbers of M allele transmitted from parents to the child on each locus as the covariates. For the quantitative trait value, we will again apply the ERS approach to obtain the upper and lower samples analogous to case and control samples.

4.2.1 Introduction to logistic model for TDT

Let $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, n$, be n independent observations from n child-parent trios, where y_i takes value 1 if the child of i-th family has the disease, or 0 otherwise and let $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{x}_i \in R^p$ contains the alleles transmission counts of the parents of i-th family at p SNP marker loci, generally, p is much larger than n, x_{ij} denote the allele transmission count of marker j and is defined as follows.

Let M and m denotes the allele names of j-th marker locus, we define

$$x_f = \begin{cases} 0 & \text{if the father is MM} \\ 0 & \text{if the father is mm} \\ 1 & \text{if the father is Mm and transmits M to the child} \\ -1 & \text{if the father is Mm and transmits m to the child.} \end{cases}$$

and

$$x_m = \begin{cases} 0 & \text{if the mother is MM} \\ 0 & \text{if the mother is mm} \\ 1 & \text{if the mother is Mm and transmits M to the child} \\ -1 & \text{if the mother is Mm and transmits m to the child.} \end{cases}$$

for father and mother, respectively. We will consider the total allele transmission counts of the father and the mother in our model, that is $x_{ij} = x_f + x_m$.

We consider the linear combination of the elements of $\mathbf{x}_i$,

$$\mathbf{x}_i^{'}\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

where $\beta = (\beta_1, \cdots, \beta_p)$ denotes the coefficient vector. Assume $\text{Prob}(y_i = 1|\mathbf{x}_i) = \pi_i$, by the logit link function

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^{'}\beta, \tag{4.4}$$

it follows therefore that the log-likelihood function is

$$l(\mathbf{y}, \beta) = \sum_{i=1}^{n} \left\{ y_i(\mathbf{x}_i^{'}\beta) - \log[1 + \exp(\mathbf{x}_i^{'}\beta)] \right\}. \tag{4.5}$$

We consider the logistic model to locate the disease susceptibility locus by allele transmission information because of an important property of the logistic model which is not shared by other link functions. This property is that no matter the data are collected prospectively or retrospectively, the logistic model can be used to make inference on the relationship between the response variable and the covariates. Our sampling scheme for dichotomous trait or for QTL mapping is in fact retrospective, that is, the individual families are sampled according to the disease status or the quantitative trait value of

the children, the cause (the transmission events) is retrospectively traced. In this sampling scheme, the response is not random, what are random are the transmission events. The logistic model specified in (4.4) and in (4.5) is formulated in the form of prospective sampling, i.e, the families are sampled according to the transmission events, and then followed for the disease status of the children. For the proof of the validness of the logistic model for both prospective and retrospective studies, we refer the reader to McCullagh and Nelder (1998).

4.2.2 LASSO and glmpath

The coefficients βs in logistic model (4.4) can be estimated by solving the following maximization problem,

$$\beta = \arg\max_{\beta} l(\mathbf{y}, \beta).$$

In genome-wide association study, the number of potential variables is much larger than the number of observations, i.e., p is much larger than n and only several of them may associate with the disease, that means most of βs should be estimated as 0. Tibshirani (1996) suggested that the above optimization problem be modified as a constrained problem:

$$\begin{aligned} &\min_{\beta} \{-l(\mathbf{y}, \beta)\} \\ &\text{s.t.} \quad \sum_{j=1}^{p} \left|\beta_j\right| \leq t, \end{aligned} \tag{4.6}$$

where $t \geq 0$ is a tuning parameter. This is a penalized likelihood problem as an extension of the LASSO. If t is greater than or equal to the l^1 norm of the ordinary maximum

likelihood estimator, then that estimator is unchanged by the LASSO. For smaller t, the LASSO shrinks the estimator towards the origin, i.e, to set some of βs equal to zero. Since there are only a small number of βs corresponding to relevant loci will have non-zero values, we need to solve the sparse optimal solutions of (4.6). The loci with non-zero βs are considered to be linked and in linkage disequilibrium with disease susceptibility loci. By the optimality conditions, it can be shown that the above constrained logistic regression is equivalent to the following unconstrained optimization problem:

$$\min_{\beta} f(\beta, \lambda) = \{-l(\mathbf{y}, \beta)\} + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{4.7}$$

where $\lambda > 0$. It is a logistic regression with L_1 penalization. There are some existing path-following approaches to estimate coefficient β of this problem. For example, Osborne et al (2000) introduce an efficient algorithm by considering the primal and dual problem; Efron et al (2004) suggested an efficient algorithm to determine the exact piecewise linear coefficient paths for LASSO. The algorithm LARS can also be used to solve this problem. Another method is Support vector machine path algorithm developed by Hastie et al (2004). In particular, we introduce L_1 regularization path algorithm namely glmpath algorithm (Park and Hastie 2006) in the following and apply it in the simulation studies.

Glmpath algorithm is a path-following procedure. Starting from $\lambda = \infty$, at each of particular values of λ, the algorithm computes the coefficient exactly, the accuracy of the path is controlled by the variant of λs. Since $f(\beta, \lambda)$ is a convex function of β, there exists a $\beta(\lambda)$ that attains the unique minimum value for each $\lambda > 0$. And, such $\beta(\lambda)$

satisfies

$$\rho(\beta, \lambda) = \frac{\delta f}{\delta \beta} = -\mathbf{x}'(\mathbf{y}-\boldsymbol{\pi})+\lambda S\, gn\binom{0}{\beta} = 0, \tag{4.8}$$

where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$ and $\pi_i = \dfrac{\exp \mathbf{x}_i'\beta}{1 + \exp \mathbf{x}_i'\beta}, i = 1, \cdots, n.$

By optimality conditions, we know that the intercept is the only nonzero coefficient when $\lambda > \max_{j\in\{1,\ldots,p\}} |\mathbf{x}'(\mathbf{y} - \bar{y}\mathbf{1})|$, and $\hat{\beta}_0 = \log(\dfrac{\bar{y}}{1-\bar{y}})$. As λ decreases, we have more and more nonzero coefficients, the index set of nonzero coefficients is named active set. At the beginning, variable $j_0 = \arg\max_j |\mathbf{x}'(\mathbf{y} - \bar{y}\mathbf{1})|$ join the active set.

In glmpath algorithm, the step length in λ, $\Delta_k = \lambda_k - \lambda_{k-1}$, was chosen to be the smallest number that will change the active set of variables. In k-th step, with the decrease in λ, the linear approximation of the corresponding change in β can be obtained by the first order expansion of $\beta(\lambda)$ at λ^k:

$$\hat{\beta}_0^{k+1} = \hat{\beta}^k + (\lambda_{k+1} - \lambda_k)\frac{\delta \beta}{\delta \lambda}.$$

To compute $\dfrac{\delta \beta}{\delta \lambda}$, for the current active set, $\rho(\beta(\lambda), \lambda)$ is zero for all λ. By differentiating $\rho(\beta(\lambda), \lambda)$ with respect to λ, it yields that

$$\frac{\delta \rho}{\delta \lambda} + \frac{\delta \rho}{\delta \beta}\frac{\delta \beta}{\delta \lambda} = \mathrm{Sgn}\begin{pmatrix}0\\ \beta\end{pmatrix} - [\mathbf{X}_\mathbf{A}'(\mathbf{Y} - \pi)\mathbf{X}_\mathbf{A}]\frac{\delta \beta}{\delta \lambda} = 0,$$

where $\mathbf{X}_\mathbf{A}'$ denotes the columns of $\mathbf{X}$ for the factors in the current active set. So we can solve $\dfrac{\delta \beta}{\delta \lambda}$ from the above equation. The above procedure is called *predictor step*.

The subsequent step is *corrector step*. After we obtain $\hat{\beta}_0^{k+1}$ in predictor step, $\hat{\beta}_0^{k+1}$ will be used as initial value for finding the exact solution β to (4.7). There are many existing algorithm to solve this differentiable convex optimization with linear constraints,

e.g Newton method and interior-point method. Since the previous predictor step has provided a warm start, that means $\hat{\beta}_0^{k+1}$ is usually close to the exact solution $\hat{\beta}^{k+1}$, the cost of solving for exact solution is relatively very low.

Thus glmpath algorithm alternates between predictor step and corrector step. After each corrector step, the glmpath algorithm has a procedure to check if the active set $\mathcal{A}$ should have been augmented. This procedure is proposed by Rosset et al (2003). For any $j \in \mathcal{A}^c$ if $\lambda < \left|\mathbf{x}_j'(\mathbf{y} - \hat{\pi})\right|$, then $\mathcal{A}$ is augmented to $\mathcal{A} \cup \{j\}$ for the next corrector step. Such procedure is repeated until the active set cannot be augmented further. Then the variables with zero coefficients are removed from the active set.

The glmpath algorithm is suitable for data consisting of much more candidate features than the sample size, because it successfully selects up to n variables regardless of the number of input features.

4.2.3 Genome-wide mapping procedure

For feature selection, we firstly use glmpath algorithm to obtain the rough sets of coefficient paths. Since the number of non-zero coefficients estimated by glmpath algorithm depends on the sample size, it tends to select too many spurious features overwhelming the causal features. To make the final feature selection, we further apply some criterion of model selection for better selection. Bayesian information criterion (BIC) is a classical dimension-consistent criterion. As pointed by Chen and Chen, BIC is too liberal

in the sense that too many spurious features will be selected when the dimension of the feature space is extremely high, it is the case in genome-wide association study. Chen and Chen (2007) recently developed EBIC particularly for feature selction with high dimension feature space. Let $\mathcal{S}$ denote the subset of integers from 1 to p, $X(\mathcal{S})$ denote the design matrix with column indices in $\mathcal{S}$ and $\beta(\mathcal{S})$ denote the corresponding components of β. The EBIC of a model with design matrix $X(\mathcal{S})$ is defined as

$$EBIC_\gamma(\mathcal{S}) = -2l(\mathbf{y}, \hat{\beta}(\mathcal{S})) + v(\mathcal{S})\ln n + 2\gamma\ln[\tau(\mathcal{S})], \tag{4.9}$$

where $\hat{\beta}(\mathcal{S})$ is the maximum likelihood estimate of $\beta(\mathcal{S})$, $v(\mathcal{S})$ is the cardinality of $\mathcal{S}$, $\tau(\mathcal{S})$ is the total number of models which can be formed by $v(\mathcal{S})$ features from the original feature space of dimension p, and γ is a constant between 0 and 1 which is to be determined by the user. The traditional BIC is a special case of EBIC with $\gamma = 0$.

Thus, for disease gene mapping or QTL mapping, we develop a generalized TDT procedure that uses the penalized logistic model and EBIC. The penalized logistic model is applied to order models and EBIC is then used to make the final selection. The detailed steps of the algorithm is described in 2 phases:

Phase I: Screening using glmpath.

In this phase, we assume logistic model (4.5) containing all alleles transmission counts of p markers. Then the penalized likelihood (4.6) is solved by glmpath algorithm. Denote the sequence of active set of each step as $S_1, \cdots, S_n$.

Phase II: Selection using EBIC.

For markers in each S_i, EBIC are computed and the set of markers in the model with the smallest values of EBIC are eventually selected.

To calculate EBIC in phase II and BIC for comparison, we use an iterative Newton-Raphson algorithm to solve the ordinary maximum likelihood estimates of (4.5) based on each of active sets obtained in phase I. However, as we have known that the existence, finiteness, and uniqueness of maximum likelihood estimates of (4.5) depend on the patterns of observed data points. The likelihood equation for a logistic regression model does not always have a finite solution. Sometimes we may have an infinite solution. This is probably resulted by the low rank of the design matrix or the separation of the data. For the response and the vector of explanatory variables of the ith subject $(y_i, \mathbf{x}_i),\ i = 1, 2, \ldots, n$, there are three types of data configurations: complete separation, quasi-complete separation, and overlap. They are defined as follows.

If there is a vector b that correctly allocates all observations to their appropriate response group, i.e

$$\begin{cases} b'\mathbf{x}_i > 0 \quad \text{for} \quad y_i = 1 \\ b'\mathbf{x}_i < 0 \quad \text{for} \quad y_i = 0. \end{cases} \qquad i = 1, 2, \ldots, n$$

then there is a *complete separation* of data points.

If there is a b such that

$$\begin{cases} b'\mathbf{x}_i \geq 0 \quad \text{for} \quad y_i = 1 \\ b'\mathbf{x}_i \leq 0 \quad \text{for} \quad y_i = 0, \end{cases} \qquad i = 1, 2, \ldots, n$$

the data is *quasi-complete separation*. When there is no separation (complete or quasi-complete) found in the data configuration, the data is *overlap*. In overlap data configuration the maximum likelihood estimates exist and are unique. But the complete or quasi-complete separation of the data may cause the maximum likelihood estimate of some βs to be infinity. Therefore, to reduce the bias in simulation studies, we apply a method of bias reduction approach in estimates of generalized linear model which was developed by Firth (1993). In logistic regression (4.4) and log likelihood (4.5), let $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ be the design matrix, then the score with respect to β is $U = \dfrac{\partial l(\mathbf{y}, \beta)}{\partial \beta}$ and the kth component of the score is

$$U_k = \sum_i (y_i - \pi_i) x_{ki}, \quad k = 1, \cdots, p.$$

Denote matrix

$$W = \text{diag}\{\pi_i(1 - \pi_i)\},$$

then Fisher information matrix for β is

$$I(\beta) = X^T W X.$$

Firth's method is indeed a penalized likelihood method, instead of maximizing the likelihood (4.5), it maximizes

$$l(\mathbf{y}, \beta) - \frac{1}{2} \ln |I(\beta)|.$$

The maximizing can be realized by adjusting y_i to $y_i + h_i/2$, where h_i is the ith diagonal element of the 'hat' matrix

$$H = W^{1/2} X I^{-1}(\beta) X^T W^{1/2}.$$

Hence, the estimated coefficients β is obtained by solving

$$U_k = \sum_i [(y_i + \frac{h_i}{2}) - (1 + h_i)\pi_i]x_{ki} = 0, \quad k = 1, \cdots, p.$$

4.2.4 Genome-wide TDT for QTL mapping

Analogous to TDT for QTL mapping at a single locus, we also apply ERS procedure to extend TDT to QTL mapping in genome-wide study. The ERS approach is as follows. For a specified integer k, each time k children from k families with parents available are chosen at random from the population. Only trait values of children are measured and ordered from the smallest to the largest. Then the family with child of rank 1 is selected as a member of upper group and the family with child of rank k is selected as a member of lower group. The genotypes of children and parents in the sample will be obtained to get the information of alleles transmission at each of the markers. The data can be summarized in Table 4.2. The upper and lower group are analogous to case and control group. Hence TDT_1, TDT_2 and TDT_3 can be employed for QTL mapping. By ERS approach, we have dichotomized the quantitative trait so that the same logic as that for dichotomous diseases applies. Therefore the logistic model (4.4) developed for disease gene mapping can be used for QTL mapping.

Table 4.2: Alleles transmission in upper group and lower group at p SNP marker loci

	No. of alleles transmitted				
	1	$\cdots$	j	$\cdots$	p
Upper	(n_M^1, n_m^1)	$\cdots$	(n_M^j, n_m^j)	$\cdots$	(n_M^p, n_m^p)
Lower	(u_M^1, u_m^1)	$\cdots$	(u_M^j, u_m^j)	$\cdots$	(u_M^p, u_m^p)
Subtotal	(t_M^1, t_m^1)	$\cdots$	(t_M^j, t_m^j)	$\cdots$	(t_M^p, t_m^p)

4.3 Numerical studies

4.3.1 Simulation setting and details for data generation

By simulation studies, we compare our generalized TDT procedure for gene mapping with classical Bonferroni-type multiple comparison approach by comparing the positive selection rate (PSR) and the false discovery rate (FDR) of each approach. Both EBIC and BIC are calculated to compare their performance in genome-wide feature selection. To make the simulation data close to reality, we use a real set of SNP markers of human being consists of 2155 markers which are distributed on 23 human chromosomes. According to the locations of the markers on a chromosome. We first calculate the genetic distance between any two marker loci. The genetic distance is measured by $Morgan(M)$ or $CentiMorgan(cM)$. The recombination fraction θ between any two adjacent marker loci on a same chromosome can be calculated by Haldane's map function: $\theta = \frac{1}{2}(1 - e^{-2d})$, where d is the genetic distance between two markers in units of Morgan.

Among these 2155 marker loci, we randomly choose 1 of 5 of them to be causal genes and assume a frequency of causal allele $f = f_1$ or $f = (f_1, f_2, f_3, f_4, f_5)$. Except these underlying loci, the frequencies of alleles of all other loci are assumed to be equally likely. The sample size is taken as $n = 400$. The haplotypes of 400 fathers and 400 mothers on 2155 loci are generated randomly according to allele frequency of each of loci. The genotypes of the offsprings are obtained by principles of human genetics.The PSR and FDR are averaged over 100 replicates of simulation. The results for disease gene mapping and for QTL mapping are depicted respectively.

To ease notation, in this subsection, we denote J the number of disease loci under consideration, namely, $J = 1$ or $J = 5$. In case-control study, according to the genotype of the child i at the disease-causing loci, the disease status of the child is determined by the total penetrance of causal loci:

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

where

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j g_{ij}, , \tag{4.10}$$

with $g_{ij} = 0, 1, 2$ corresponding to the genotype of jth causal locus being mm, Mm, MM, respectively $j = 1, \ldots, J$. The response y_i is either 1 (affected) or 0 (unaffected) with the probability π_i and $1 - \pi_i$.

For each simulated data set, the three TDT tests TDT_1, TDT_2 and TDT_3 based on case group only, case and control group and control group only are conducted on each

of marker loci. Markers with highest significant effects are selected with false discovery rate controlling (FDRC) procedure with the critical level 0.05.

In the simulation of QTL mapping study, we assume that effects of the QTL alleles are additive. The quantitative trait value of the offspring is generated according to $X = g + \epsilon$, where g represents the expected genotypic values determined by the genotype of causal loci and ϵ is a random variable with the standard normal distribution. The batch size in ERS, k, is taken to be 10 and 20. Each replicate of ERS samples is generated as follows. First, 10 or 20 copies of families with one child and two parents are generated independently. The quantitative trait value X of the child and the genotypes of all family members are known. Then, the families are ranked with respect to X, after that, the family with smallest rank is put into the lower sample which is similar to the control group and the family with the largest rank is put into the upper sample which is similar to the case group in dichotomous case-control study.

For the families in upper and lower samples, we can calculate the alleles transmission counts at all of the marker loci from the genotypes of parents and child. Then the TDT_1, TDT_2 and TDT_3 are conducted on upper and lower sample at each marker locus and the respective p-value is obtained. This procedure is then followed by FDRC approach. The locus which is judged significant by FDRC procedure are selected. On the other hand, the logistic model (4.5) is assumed with 2155 covariates by treating the upper and lower samples as case and control samples. That means the response is assumed 1 for families in the upper sample and 0 for families in the lower sample. Then

we apply generalized TDT cum EBIC introduced in last section to select markers which have genetic effect on the quantitative trait. BIC is also calculated for comparison. In calculation of EBIC and BIC, the corresponding maximum likelihood estimations are obtained by the approach developed by Firth (1993). Then the set of markers which give the lowest EBIC or BIC are selected and considered to regulate the quantitative trait or have linkage disequilibrium with QTL.

4.3.2 Simulation results for case-control study

The logistic model (4.4) is assumed with 2155 covariates. Then glmpath algorithm is applied to get the ordered sets of feature, with each of these sets of features, the ordinary un-penalized likelihood (4.5) is maximized and BIC and EBIC are calculated respectively. This procedure is denoted by LOGIT-BIC and LOGIT-EBIC. The results under different settings are given in Table 4.3 and Table 4.4. In these tables, the cases (i)–(vi) represent different parameter settings for β in (4.10)and causal allele frequency f. These values of β and f are given as follows:

From Table 4.3, it can be seen that PSR and FDR of TDT_1, TDT_2 and TDT_3 with false discovery rate controlling approach varies with different causal allele frequencies. For the very common disease that means causal alleles have relatively large frequencies and penetrance (big value of β), e.g, case ii in the table, TDT_3 has the best performance (PSR is 0.748 which is comparable to PSR of TDT_2 with smallest FDR 0.051) whereas TDT_1 has very small PSR (0.012) and very big FDR (0.667). It is a rare case from

Case i	$\beta = (-6, 1.8, 1.5, 1.3, 1.1, 0.9)$	and	$f = (0.3, 0.32, 0.35, 0.38, 0.42)$,
Case ii	$\beta = (-6, 2.8, 2.5, 2.3, 2.1, 1.9)$	and	$f = (0.3, 0.32, 0.35, 0.38, 0.4)$,
Case iii	$\beta = (-6, 2.8, 2.5, 2.3, 2.1, 1.9)$	and	$f = (0.2, 0.22, 0.25, 0.28, 0.3)$,
Case iv	$\beta = (-6, 2.8, 2.5, 2.3, 2.1, 1.9)$	and	$f = (0.1, 0.12, 0.15, 0.18, 0.2)$,
Case v	$\beta = (-6, 2.8, 2.5, 2.3, 2.1, 1.9)$	and	$f = (0.04, 0.05, 0.055, 0.06, 0.07)$.
Case vi	$\beta = (-6, 1.7)$	and	$f = 0.35$.

the evolution genetics point of view unless allelic drift is strong in small populations or strong interaction between genotype and environment exists. These three TDTs are all sensitive to the change of allele frequency or penetrance. In case i, with the same allele frequencies as case ii, we change penetrance of the alleles to be smaller, a big difference is found in the result that TDT_1 and TDT_2 have better PSR (0.456 and 0.452) than TDT_3 (0.006). With the fixed penetrance, each causal allele frequency varies from bigger to smaller, we found that TDT_1 can give us the better and better selection; in contrast with TDT_1, TDT_2 performs worse and worse, even TDT_3 doesn't work at all finally (e.g case iv and case v).

From Table 4.4, it is verified that BIC tend to select too many spurious variables compared with EBIC, this is indicated by the very high PSR and very high FDR as well. Thus, it is shown that BIC is not suitable for feature space of high dimension. In contrast with TDTs cum FDRC procedure, our generalized TDT cum EBIC is not so sensitive to the allele frequency and penetrance as TDTs cum FDRC does, especially to allele frequency which is usually unknown in practice. This superiority is illustrated in

case ii – v, with the same penetrance, our TDT cum EBIC have almost stable PSR and FDR with various allele frequencies whereas TDTs cum FDRC have big fluctuation. In particular, by EBIC with $\gamma = 1$ we can always obtain very low FDR.

Table 4.3: PSR and FDR of various TDTs with FDRC for disease gene mapping

	PSR			FDR		
Cases	TDT_1-FDRC	TDT_2-FDRC	TDT_3-FDRC	TDT_1-FDRC	TDT_2-FDRC	TDT_3-FDRC
i	0.456	0.452	0.006	0.054	0.051	0.667
ii	0.012	0.79	0.748	0.455	0.057	0.041
iii	0.228	0.704	0.122	0.088	0.054	0.141
iv	0.848	0.634	0	0.058	0.068	1
v	0.99	0.464	0	0.037	0.076	1
vi	1	1	0	0.083	0.057	1

Table 4.4: PSR and FDR of logistic regression cum BIC and EBIC for disease gene mapping

	PSR			FDR		
Cases	BIC	EBIC($\gamma = 0.5$)	EBIC($\gamma = 1$)	BIC	EBIC($\gamma = 0.5$)	EBIC($\gamma = 1$)
i	0.88	0.638	0.388	0.805	0.229	0.035
ii	0.982	0.866	0.644	0.783	0.198	0.009
iii	0.984	0.9	0.676	0.791	0.174	0.023
iv	0.992	0.89	0.692	0.776	0.155	0.019
v	0.992	0.8	0.604	0.772	0.128	0
vi	1	0.96	1	0.953	0.178	0.038

4.3.3 Simulation results for QTL mapping

From Table 4.5 and Table (4.6), we can see that the feature selection approaches original for case-control study can perform very well by dichotomizing the quantitative trait with ERS procedure. The simulation results are obtained under different heritability of the quantitative trait since the allele frequencies and allele effect are compounded in heritability. We performe extensive investigations in the parameter space where the heritability varies from 0.05 around to 0.5 around, and $\gamma = 0.5, 1$ in EBIC. It is noted in Table 4.5 and Table (4.6) that, for the increasing heritability, every approach has better performance. However, compared with other methods, BIC always has highest PSR but severely high FDR, this is a crucial deficiency in practice. Therefore, BIC is more likely not suitable for variable selection in high dimension space, since it tends to select too

many spurious variables. It is worth to see that with comparable PSR our generalized TDT cum EBIC always gives lowest FDR when we take $\gamma = 1$. We have found that in disease susceptibility locus mapping, TDT_2 is better than TDT_1 and TDT_3 for common disease, but for rare disease TDT_1 preforms best. In contrast, in QTL mapping with ERS, TDT_2 is almost better than TDT_1 and TDT_3 for various heritabilities in the sense that it has higher PSR and comparable FDR. This is because we assumed relatively larger frequencies of SNP maker alleles in QTL mapping. These simulation data is generated with the ERS batch size k being 10 and 20. It is expected that all the methods can perform better with the batch sizes $k = 20$ than $k = 10$. The results in Table 4.7 and Table 4.8 verified this fact. However, in their QTL mapping studies, Lander and Botstein (1989) warned that very extreme trait values might have causes other than genetic effects. They suggested that for the truncation approach the upper and the lower percentage should not be less than 5%. That means the batch size k should not be more than 20.

Table 4.5: PSR and FDR of various TDTs with FDRC for QTL mapping (ERS batch size $k = 10$)

	PSR			FDR		
h	TDT_1-FDRC	TDT_2-FDRC	TDT_3-FDRC	TDT_1-FDRC	TDT_2-FDRC	TDT_3-FDRC
0.056	0	0.034	0.004	1	0.15	0.6
0.168	0.148	0.4	0.148	0.109	0.104	0.149
0.308	0.422	0.884	0.384	0.079	0.062	0.06
0.462	0.736	0.982	0.77	0.062	0.063	0.075
0.570	0.892	0.998	0.896	0.090	0.035	0.063

Table 4.6: PSR and FDR of generalized TDT cum BIC and EBIC for QTL mapping (ERS batch size $k = 10$)

	PSR			FDR		
h	BIC	EBIC($\gamma = 0.5$)	EBIC($\gamma = 1$)	BIC	EBIC($\gamma = 0.5$)	EBIC($\gamma = 1$)
0.056	0.354	0.109	0.097	0.919	0.667	0.586
0.168	0.7	0.484	0.332	0.852	0.277	0.067
0.308	0.98	0.896	0.714	0.790	0.177	0.013
0.462	0.998	0.98	0.92	0.74	0.162	0.025
0.570	0.998	0.996	0.974	0.756	0.156	0.024

Table 4.7: PSR and FDR of various TDTs with FDRC for QTL mapping (ERS batch size $k = 20$)

	PSR			FDR		
h	TDT_1-FDRC	TDT_2-FDRC	TDT_3-FDRC	TDT_1-FDRC	TDT_2-FDRC	TDT_3-FDRC
0.056	0.046	0.168	0.028	0.258	0.116	0.44
0.168	0.274	0.534	0.258	0.127	0.076	0.092
0.308	0.73	0.986	0.674	0.069	0.060	0.051
0.462	0.926	1	0.944	0.080	0.042	0.050
0.570	0.99	1	0.994	0.063	0.060	0.054

Table 4.8: PSR and FDR of generalized TDT cum BIC and EBIC for QTL mapping (ERS batch size $k = 20$)

	PSR			FDR		
h	BIC	EBIC($\gamma = 0.5$)	EBIC($\gamma = 1$)	BIC	EBIC($\gamma = 0.5$)	EBIC($\gamma = 1$)
0.056	0.528	0.258	0.18	0.890	0.466	0.25
0.168	0.804	0.582	0.416	0.818	0.261	0.041
0.308	0.998	0.968	0.938	0.770	0.175	0.025
0.462	1	0.996	0.95	0.733	0.124	0.010
0.570	1	1	0.99	0.70	0.155	0.018

4.4 A new algorithm for logistic model with grouped variables

4.4.1 Penalized logistic model with grouped covariates

In the above sections, TDT is generalized to genome-wide association study by using a logistic model with the covariates being the sum of parents' allele transmission counts. The two parents of a family are independent, so the transmission counts of them can be used in the regression model separately. In this section, we reconsider the feature selection problem by using these separate transmission counts and consider a group LASSO algorithm.

Consider jth SNP marker locus with two alleles, namely M and m. For a parent-child trio, let x_{fj} and x_{mj} denote the value of allele transmission count at this marker locus for father and mother respectively. They are defined as follows:

$$x_{fj} = \begin{cases} 0 & \text{if the father is MM} \\ 0 & \text{if the father is mm} \\ 1 & \text{if the father is Mm and transmits M to the child} \\ -1 & \text{if the father is Mm and transmits m to the child.} \end{cases}$$

and

$$x_{mj} = \begin{cases} 0 & \text{if the mother is MM} \\ 0 & \text{if the mother is mm} \\ 1 & \text{if the mother is Mm and transmits M to the child} \\ -1 & \text{if the mother is Mm and transmits m to the child.} \end{cases}$$

Then the j-th factor X_j is $X_j = (x_{fj}, x_{mj})$. Assume that

$$\text{Prob}(y_i|X_{i1}, \ldots, X_{iJ}) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

and $y_i = 1$ or -1 means the child is affected or not affected. Given the data of parent-child trios (y_i, x_{fj}, x_{mj}), by the logit link function, we have

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{J} X_{ij}'\beta_j,$$

where $\beta_j = (\beta_{fj}, \beta_{mj})'$ is a coefficient vector of size 2. The log-likelihood function is

$$-\sum_{i=1}^{n} \log\left(1 + \exp\left\{-y_i(\beta_0 + \sum_{j=1}^{J} \beta_j' X_{ij})\right\}\right). \tag{4.11}$$

Using the idea of group LASSO, the relevant marker loci can be selected by solving the sparse logistic regression optimization problem with grouped variables:

$$\min_{\beta} \quad l(\beta) = \sum_{i=1}^{n} \log\left(1 + \exp\left\{-y_i(\beta_0 + \sum_{j=1}^{J} \beta_j' X_{ij})\right\}\right) + \lambda \sum_{j=1}^{J} \|\beta_j\|, \quad \lambda > 0, \tag{4.12}$$

where $\|\beta_j\| = \sqrt{\beta_{fj}^2 + \beta_{mj}^2}$. Since $\|\beta_j\|$ is a convex function on R^2 for each $j = 1, \ldots, J$, so it follows that $\lambda \sum_{j=1}^{J} \|\beta_j\|$ is convex with respect to β on R^{2J} for any $\lambda > 0$. In addition, note that

$$\log\left(1 + \exp\left\{-y_i(\beta_0 + \sum_{j=1}^{J} \beta_j' X_{ij})\right\}\right)$$

is also a convex function in β. Thus, the objective function $l(\beta)$ in problem (4.12) is convex on R^{1+2J}.

Hence, by the theory of nonlinear programming (NLP), the optimality conditions for minimizing $l(\beta)$ over R^{1+2J} are necessary and sufficient conditions. So, we can derive the first order optimality conditions for the unconstrained convex problem (4.12). Note that $l(\beta)$ is differentiable with respect to β_0 and is differentiable in β_j if $\|\beta_j\| \neq 0$, $j = 1, 2, \ldots, J$, and $l(\beta)$ is nondifferentiable in β_j at the points where $\|\beta_j\| = 0$, $j = 1, 2, \ldots, J$. For the latter case, the notion of subdifferential will be introduced for characterizing the optimality conditions for this convex problem.

Now, suppose that $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_J^*)' \in R^{1+2J}$ with $\beta_j^* = (\beta_{fj}^*, \beta_{mj}^*)'$ is an optimal solution of problem (4.12). Then, the optimality conditions for (4.12) can be written as follows.

$$\begin{cases} (i) & \left.\dfrac{\partial l(\beta)}{\partial \beta_0}\right|_{\beta=\beta^*} = 0, \\ (ii) & \text{For } j \in \{1, 2 \ldots, J\}, \quad \text{if} \quad \|\beta_j^*\| \neq 0, \quad \text{then} \\ & \left.\nabla_{\beta_j} l(\beta)\right|_{\beta=\beta^*} = 0, \\ (iii) & \text{For } j \in \{1, 2 \ldots, J\}, \quad \text{if} \quad \|\beta_j^*\| = 0, \quad \text{then} \\ & 0 \in \left.\partial_{\beta_j} l(\beta)\right|_{\beta=\beta^*}, \end{cases} \tag{4.13}$$

where $\nabla_{\beta_j} l(\beta)$ denotes the gradients of l with respect to the subvector β_j. Similarly, $\partial_{\beta_j} l(\beta)$ denotes the subdifferential of l with respect to β_j. Recall that the notion of subdifferential, $\partial H(x)$, at x of a convex real valued function $H : R^n \to R$ is defined as

follows.

$$\partial H(x) := \{w \in R^n \mid H(z) \geq H(x) + \langle w, z - x\rangle, \text{ for all } z \in R^n\}.$$

Note that $\partial H : R^n \to R^n$ is a set-valued mapping and $\partial H(x)$ is a nonempty convex set for any $x \in R^n$. In particular, $\partial H(x)$ is reduced to a singleton $\{\nabla H(x)\}$ if H is differentiable at x. For instance, consider the absolute value function $f : R \to R$, defined by $f(x) = |x|$. Clearly, $f(x)$ is a convex function on R and is differentiable everywhere except at the origin. In this case, by definition, $\partial f(x) = \{1\}$ if $x > 0$, $\partial f(x) = \{-1\}$ if $x < 0$, and the subdifferential of f at the origin is a closed interval, i.e.,

$$\partial f(0) = [-1, 1].$$

To ease the notation, we define

$$\begin{aligned}
\eta_i &= -y_i(\beta_0 + \sum_{j=1}^{J} \beta_j' X_{ij}), \\
\Gamma_0 &= \sum_{i=1}^{n} \frac{e^{\eta_i}}{1 + e^{\eta_i}} y_i, \\
\Gamma_j &= \sum_{i=1}^{n} \frac{e^{\eta_i}}{1 + e^{\eta_i}} y_i X_{ij}, \quad j = 1, 2, \ldots, J.
\end{aligned}$$

Then, the optimality conditions (4.12) can be rewritten as follows:

$$\begin{aligned}
&\Gamma_0 = 0, \text{ if } j = 0; \\
&\Gamma_j - \lambda \frac{\beta_j}{||\beta_j||} = 0, \text{ if } j \in \{1, 2, \ldots, J\} \text{ and } ||\beta_j|| \neq 0; \\
&||\Gamma_j|| \leq \lambda, \text{ if } j \in \{1, 2, \ldots, J\} \text{ and } ||\beta_j|| = 0.
\end{aligned}$$

Define viol_j, $j = 1, 0, 1, \ldots, J$, by

$$\text{viol}_j := \begin{cases} |\Gamma_0| & \text{if } j = 0 \\ \left\| \Gamma_j - \lambda \frac{\beta_j}{\|\beta_j\|} \right\| & \text{if } j > 0 \text{ with } \|\beta_j\| \neq 0 \\ \max\{0, \|\Gamma_j\| - \lambda\} & \text{if } j > 0 \text{ with } \|\beta_j\| = 0. \end{cases} \tag{4.14}$$

Then the optimality conditions of (4.12) can be further written as

$$\text{viol}_j = 0, \quad \text{for any } j.$$

Thus, to derive the optimal solution of (4.12), we only need to find the solution of the system of the above equalities. Unfortunately, an obstacle for solving this system is that in practice it is hard or even impossible to achieve the exact optimality conditions in finite time. To overcome this, we turn to relax the above optimality conditions to a system of inequalities by introducing a tolerance ν, which is a very small positive number. That is,

$$\text{viol}_j \leq \nu, \quad \text{for any } j.$$

Then, the algorithm will stop when the above system of inequalities holds for a designated tolerance $\nu > 0$. Note that the obtained solution is called ν-optimal to the original problem (4.12) in NLP.

4.4.2 An algorithm for variable selection

According to the optimality conditions derived in last section, we provide an algorithm for solving the optimization problem (4.12) efficiently. This algorithm is analogous to

the algorithm proposed by Shevade and Keerth (2003). We obtain the sparse group solution by a two-loop approach. For a given β, the index set $I = \{0, 1, 2, \ldots, J\}$ can be divided into two subsets of indices according to whether the subvector β_j is zero or not:

$$I = I_0 \cup I_1$$

$$I_0 = \{j : \|\beta_j\| = 0, j > 0\}$$

$$I_1 = \{0\} \cup \{j : \|\beta_j\| \neq 0, j > 0\}.$$

For a fixed value of λ, all βs are set to be zero initially. We summarize the algorithm as follows:

- **Phase I:**

  Calculate viol_j for current β and choose the maximum violator, m, in I_0, i.e.,

  $$m = \arg\max_j \{\text{viol}_j \mid j \in I_0\}.$$

  Let $I_1 = \{m\} \cup I_1$, go to phase II.

- **Phase II:**

  Solve the optimization subproblem $\min l(\beta)$ with respect to β_{I_1}, go back to Phase I.

For a sufficient large $\lambda > 0$, from the definition of Γ, we can see that the intercept β_0 is the only nonzero coefficient. As λ decreased, there are some violators in $\text{viol}_j = 0$, $j = 1, \ldots, J$, for the current solution. The variable with the index $m = \arg\max_j \text{viol}_j$ is selected and the coefficient β_m is estimated by solving the optimization problem (4.12) with respect to β_{I_1}.

Chapter 5

Conclusion and further research

5.1 Conclusion

TDT is a powerful method of LD mapping, it is applied to map qualitative trait locus first. Many researchers have extend it to QTL mapping, for example t-test with random sampling requiring exactly one heterozygous and one homozygous parent, and χ^2 test with truncation sampling. The truncation sampling needs a pre-screening process to estimate the cutoff quantiles if we don't know them a priori, this pre-screening process can be very time-consuming and incur a nonnegligible cost and unnecessary errors. In the pre-screening process some individuals may be lost to follow-up and so forth. In this thesis, we provide an alternative ERS approach to extend TDT to QTL mapping. With the comparable power, the TDT with ERS have the advantage of easy implementation. It does not require a pre-screening process. The selection is done in batches of a small

number of k individuals and is well within the manageable range. In the simulation study, we compared the power of various TDT statistics under different sampling approaches. We conclude that although TDT with truncation sampling scheme is slightly more powerful than TDT with ERS approach, it is more difficult to implement in certain situations than is the TDT with ERS approach for QTL mapping because truncation scheme requires a process of presceening which is usually not simple in practice. A large amount of individuals are required for estimating the cutoff quantiles of a random variable. Generally, it is not simple to keep the records of the individuals involved in the prescreening process and to recall them for genotyping usually after a long period. Spielman et al.(1993) suggested that we should consider TDT incorporating unaffected offspring when we have concern of segregation distortion. Deng and Chen (2001) compare the power of original TDT and the TDT with unaffected offsprings involved, they found that with larger genetic effect or lager prevalence of disease, or larger frequency of disease allele, the TDT with case-control group is more powerful than others. Consistent with this conclusion, we found that with the increasing heritability of the QTL, the power of our TDT_u, TDT_l and TDT_{ul} with ERS all have increasing power, and under the small frequency of increasing allele, TDT_u is slight more powerful than the other two.

Genome-wide association study is a hot area. Hence not only we extended TDT to QTL mapping by ERS sampling approach in single locus, also we extend TDT to genom-wide disease susceptibility gene mapping and QTL mapping. In this thesis, TDT is, in the first time, applied in genome-wide association study by our general-

ized TDT cum EBIC to search the disease susceptibility gene and QTL mapping. In our approach, we construct a logistic model with allele transmission values of parents in the family-trios being covariates, since the number of SNP markers is usually quite large, the classic method of estimation is not suitable. Thus, in our thesis we provided a two-step algorithm to obtain the sparse solution of this model in high dimension space. In the gene mapping point of view, a sparse solution corresponds to the genetic loci on chromosome. In our approach, the first step is the rough selection in which glm-path algorithm (Park and Hastie, 2006)is carried out to find the solution path of the logistic model. After that, we obtain the ordered sets of locus which may regulate the quantitative trait or relate to the disease. In the subsequent refining step, a new variable selection criterion EBIC is applied for further selection. Our approach has the following advantages: (i) It is robust to the frequencies of causal alleles. In the simulation study for disease susceptibility gene mapping, we compare our approach with classic multiple-comparison approach in which TDT is performed at each of locus then the false discovery rate controlling (FDRC) procedure is applied. We found that the performance of FDRC can be very fluctuant with various causal allele frequencies, in contrast our penalized logistic model cum EBIC approach is very robust to common disease and rare disease, this is meaningful in practice since the allele frequency is usually unknown. (ii) With the various choices of γ, it can provide lower FDR compared with FDRC. In the simulation study, we find that although BIC with the same penalized logistic model provide very high PSR, the FDR of it is also too large to be acceptable. On the contrast, EBIC almost always has quite lower FDR but with comparable PSR.

(iii) It is easy to take into account the interaction of genetic and environmental effect. As we have known that there may exist strong interaction of gene and environment in human being or in animals, for example Valdar et.al (2006) showed that environmental and physiological covariates are involved in an unexpectedly large number of significant interactions with genetic background in their study of gene function using mouse model. Kraft et al. (2007) exploit the genetic and environmental interactions in their gene association study. In our logistic model, we can consider the interaction between genetic and environmental effects by adding an additional factor. (iv) This last advantage comes from the logistic function itself. That is the differences on the logistic scale can be estimated regardless of whether the data are sampled prospectively or retrospectively. In other words, although our sampling scheme is retrospective i.e the subjects involved in the study are often hospital records collected over a period of long time, the logistic model for prospective study still can be applied.

In QTL mapping in genome-wide study, we verified again that ERS is efficient for multiple-comparison of TDT approach and generalized TDT with EBIC as well. By applying ERS to dichotomize a quantitative trait, we are able to extend TDT to QTL mapping in genome wide. From the simulation result, we found that with various heritabilities of the quantitative trait, our ERS cum TDT with EBIC can achieve comparable PSR whereas lowest FDR especially with $\gamma = 1$. Therefore, it is expected that with the above advantages our method can be applied in practice to search for genes which is meaningful in genetic diagnosis and new drug development.

In this thesis, according to particularity of TDT family data, we provided a logistic model with grouped covariates which contains mother and father's allele transmission value separately. The advantage of considering the parents separately is that a locus is selected as long as the effect of one of the parents is significant. By summing up the transmission values of parents, some difference may be neglected. For example, the parents with the value $(1,-1)$ or $(-1,1)$ are considered having the same effects with the parents with the value $(0,0)$. Another advantage is that we are able to take into account gender effect by considering paternal and maternal effects separately. There are some existing algorithms to solve the penalized likelihood problem with grouped variables, for example, group LASSO (Yuan and Lin 2006). By determining some optimality conditions of the corresponding optimization program, we derived an efficient algorithm for sparse solutions of this optimization program.

5.2 Further research topics

In the various TDT approaches, it is noted that only information of allele transmission is applied whereas the exact genotypes of the parents and children are neglect, so if we combine the transmission information and the children's genotypes, not only the power of detecting QTL is expected to be improved but also the population stratification does not distort the result. On the other hand, in applying ERS sampling approach or truncation approach on TDT, the power is affected by sample size n and the batch size k and τth quantile and the k and τth quantile are predetermined by the user. The smaller

the τ or the larger the k, the more powerful the tests. However, we can not choose the very small values of τ or very big value of k because the very extreme value of the quantitative values may result from non-genetic effects. For a particularly required power, how to determine desired sample size n, batch size k and τth quantile requires further study. On the other hand, if we consider the two main costs in TDT tests which are screening cost and genotyping cost, how to gain the maximum power with constraint of the total cost is worthy of further research.

In our generalized TDT with logistic model cum EBIC. It is apparent that with the higher value of γ in EBIC, we can obtain lower FDR, but meanwhile the PSR is also lower. How to determine the γ to balance PSR and FDR requires further study. In disease gene mapping problem, we found that with the fixed genetic effects, the allele frequencies have big effect on the performance of these methods, especially on TDT_1, TDT_2 and TDT_3 with FDRC, but the effect on our method is not so distinct. We will investigate the reason of that and explore the explicit relation between these parameters and FDR and PSR of our methods in our future work.

In addition, it is known that complex disease results from the interplay of genetic and environmental factors. However, we are currently unclear how gene-environment interaction can best be used to locate complex disease susceptibility loci, particularly when large amount of markers are scanned for association with disease. We will consider this issue with our generalized TDT cum EBIC method, and some other possible tools rather than TDT may join to test association and interaction of genetic and envi-

ronmental effects as well.

Variable selection in high dimension space is a general issue in genome-wide association study. I have derived some optimality conditions of penalized likelihood function with grouped variables and an algorithm was provided for a special case of the covariates, i.e, each of the groups of covariates contains two variables. We will do some further studies on this problem in more general situations and the numerical studies are required for the real genetic data in practice.

References

Abecasis, G. R., Cardon, L. R. & Cookson, O. C. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* **66**, 279–292.

Allison, D. B. (1997). Transmission-disequilibrium test for quantitative traits. *American Journal of Human Genetics* **60**, 676–690.

Benjamini, Yoav. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, Statistical methodology* **57**, 289–300.

Betensky, R. A. & Rabinowitz, D. (2000). Simple approximation for the maximal transmission/ disequilibrium test with a multi-allelic marker. *Annals of Human Genetics*, **64**, 567–574.

Bickeboller, H. & Clerget-Darpoux, F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genetic Epidemiology* **12**, 865–870.

Bink, M. C. A. M., Te Pas, M. F. W., Harders, F. L. & Janss, L. L. G. (2000) A transmission/disequilibrium test approach to screen for quantitative trait loci in two selected lines of large white pigs. *Genetical Research* **75**, 115–121.

Chen, J. & Chen, Z. (2008). Extend Bayesan Information Criteria for Model Selection with Large Model Space. *Biometrika*, To appear.

Chen, Z. & Chen, J. (2007). Tournament screening cum EBIC for feature selection with high dimensional feature space. *report*.

Chen, Z., Zheng, G. Ghosh, K. & Li, Z. (2005). Linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *American Journal of Human Genetics* **77**, 661–669.

Cleves, M. A., Olson, J. M. & Jacobs, K. B. (1997). Exact transmission-disequilibrium tests with multiallelic markers. *Genetic Epidemiology* **14**, 337–347.

Deng, H. & Li, J. (2002). The effects of selected sampling on the transmission disequilibrium test of a quantitative trait locus. *Genetical Research* **79**, 161–174.

Deng H. W. & Chen, W. M. (2001). The power of the transmission disequilibrium test (TDT) with both case-parent and control-parent trios. *Genetical Research* **78**, 289–302.

Dudbridge, F., Koeleman, P. C., Todd, J. A. & Clayton, D. G. (2000). Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *American Journal of Human Genetics* **66**, 2009–2012.

Dudoit, S., Speed, T. P. (2000) A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. *Biostatistics* **1**, 1–26.

Duffy. D. L. (1995). Sceening a 2 cM genetic map for allelic association: A simulated oligogenic trait. *Genetic Epidemiology* **12**, 595–600.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Ewens, W. J. & Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics* **57**, 455–464.

Fan, J., & Liu, R. (1999). Variable Selection via Penalized Likelihood. *eScholarship Repository, University of California.* http://repository.cdlib.org/uclastat/papers/

Fan, J. & Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of American Statistical Association* **96**, 1348–1360.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

George, V., Tiwari, H. K., Zhu, X. & Elston, R. C. (1999). A test of transmission/ disequilibrium for quantitative traits in pedigree data, by multiple regression. *The American Journal of Human Genetics* **65**, 236–245.

Haseman, J. K., Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.

Hauser, E.R., Boehnke, M. (1998) Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics* **54**, 1238–1246.

Huang, J., Ma, S., Xie, H. & Zhang, C.(2007). A group bridge approach for variable selection. *report*.

Hunter, D. & Li, R. (2005). variable selection via MM algorithms. *Annals of Statistics* **33**, 1617–1642.

Ishwaran, H. & Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of American Statistical Association* **98**, 438- 455.

Jorde, L. B. (1995). Linkage disequilibrium as a gene-mapping tool. *American Journal of Human Genetics* **56**, 11–14.

Knapp, M. (1999). The Transmission/Disequilibrium Test and Parental-Genotype Reconstruction: The Reconstruction-Combined Transmission/ Disequilibrium Test. *American Journal of Human Genetics* **64**, 861–870.

Koeleman, B. P. C., Dubridge, F., Cordell, H. J. & Todd J. A. (2000). Adaption of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the conditional extended transmission/disequilibrium test. *Anals of Human Genetics* **64**, 207–213.

Kong, A., Cox, N. J. (1997). Allele-sharing models: lod scores and accurate linkage tests. *The American Journal of Human Genetics* **61**, 1179–1188.

Kraft, P., Yen, Y., Stram, D. O., Morrison, J., Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63**,111–119.

Lander, E. S., Botstein, D. (1989). Mappin Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Li, Y., Campbell, C. & Tipping, M. (2002). Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics* **18**, 1332–1339.

Marchini, J., Donnelly, P. & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417.

Marchini, J., Myers, S., McVean, G. & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.

Martin, E. R., Bass, M. P. & Kaplan N. L. (2001). Correcting for a Potential Bias in the Pedigree Disequilibrium Test *American Journal of Human Genetics*. **66** 1065–1067.

McCullagh, P. & Nelder, J. A. (1998). Generalized linear models. 2nd edition, Chapman and Hall.

Monks, S. A. Kaplan, N. L. (2000). Removing the Sampling Restrictions from Family-Based Tests of Association for a Quantitative-Trait Locus. *American Journal of Human Genetics* **66**, 576–592.

Morton, N. E. (1995). Significance levels in complex inheritance. *The American Journal of Human Genetics* **7**, 277–318.

Nagelkerke, N. J., Hoebee, B., Teunis, P. & Kimman, G. (2004). Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *European Journal of Human Genetics* **12**, 964–970.

Nicodemus, K. K., Luna, A. & Shugart Y. Y. (2007). An Evaluation of Power and Type I Error of Single-Nucleotide Polymorphism Transmission/DisequilibriumCBased Statistical Methods under Different Family Structures, Missing Parental Data, and Population Stratification. *The American Journal of Human Genetics* **80**, 178–175

Olson, J. M., Witte, J. S. & Elston, R. C. (1999). Tutorial in biostatistics genetic mapping of complex traits. *Statistics in Medicine* **18**, 2961–2981.

Osbornel, M. R., Presnell, B. & turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–377.

Ott, J. (1991). Analysis of human genetic linkage, *Johns Hopkins University Press, Baltimore, MD*.

Page, G. P. & Amos. C. I. (1999). Comparison of linkage-disequilibrium methods for localization of genes influencing quantitative traits in humans. *American Journal of Human Genetics* **64**, 1194–1205.

Park, M. Y. & Hastie, T. (2006). L_1 Regularization Path Algorithm for Generalized Linear Models. *report*

Pritchard, J. K., Stephens, M., Rosenberg, N. A.,& Donnelly, P. (2000). Association Mapping in Structured Populations. *American Journal of Human Genetics* **66**, 605–614.

Rabinonwitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Human Heredity* **47**, 342–350.

Rebai, A., Goffinet, B., Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics* **51**, 87–99.

Risch, N, Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589.

Satagopan, J. M., Yandell, B. S., Newton, M. A., Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. *Genetics* **144**, 805–816.

Schaid, D. J. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology* **13**, 423–450.

Schwarz, G. (1978). Estimating the dimension of a Model. *The Annals of Statistics* **6**, 461–464.

Sebastiani, P., Abad, M. M., Alpargu, G. & Ramoni, M. F. (1999). Robust transmission/disquelibrium test for oncomplete family genotypes. *Genetics* **168**, 2329–2337.

Shevade, S. K. & Keerth, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19**, 2246–2253.

Sham, P. C., Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analysis for sib pairs. *The American Journal of Human Genetics* **68**, 1527–1532.

Sham, P. C., Purcell, S., Cherny, S. S., Abecasis, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics* **71**, 238–253.

Sham, P. C. & Curtis, D. (1995). An extend transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of human genetics* **59**, 323–336.

Sinsheimer, J. S., Blangero, J. & Lange, K. (2000). Gamete-competition models. *American Journal of Human Genetics* **66**, 1168–1172.

Slatkin, M. (1999). Disequilibrium mapping of a quantitative-trait locus in an expanding population. *American Journal of Human Genetics* **64**, 1765–1773.

Spielman, R. S. & Ewens, W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**, 450–458.

Spielman, R. S. & Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59**, 983–989.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1992). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus(IDDM). *American Journal of Human Genetics* **52**, 506–516.

Sun, F. Z., Flanders, W. D., Yang, Q. H. & Zhao, H. Y. (2000). Transmission/disequilibrium tests for quantitative traits. *Annals of human genetics* **64**, 555–565.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine* **16** 385–395.

Tibshirani, R. (1996). Regression Shrinkage and selection via the lasso *Journal of the Royal Statistical Society. Series B, Statistical methodology.* **58**, 267–288.

Valdar, W., Solberg, L.C., Gauguier, D., Cookson, W. O., Rawlins, J. N. P., Mott, R., Flint, J. (2006). Genetic and environmental effects on complex traits in mice. *Genetics* **174**, 959–984.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd J. A. (2005). Genome-wide association studies: theoretical and practical concerns, K. (2002). *Nature* **6** 109–118. Fine-Scale

Xiong. M. M., Guo. S (1997). Fine-Scale Genetic Mapping Based on Linkage Disequilibrium: Theory and Applications . *The American Journal of Human Genetics* **60**, 1513-1531.

Xiong, M. M., Krushkal, J. & Boerwinkle, E. (1998). TDT statistics for mapping quantitative trait loci. *Annals of Human Genetics* **62**, 431–452.

Xu, X., Weiss, S., Xu, X., Wei, L. J. (2000). A unified Haseman-Elston method for testing linkage with quantitative traits. *The American Journal of Human Genetics* **67**, 1025–1028.

X, Z., Kerstann, K. f., Sherman, S. L., Chakravari, A. & Feingold, E. (2004). A trisomic transmission disequilibrium test. *Genetic Epidemiology* **26**, 125–131.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika* **92**, 937–950.

Yuan, Y. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **68**, 49–67.

Zeng, Z-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Zhao, J., Boerwinkle, E. & Xiong, M. M. (2007). An entropy-based genome-wide transmission/disequilibrium test. *Human Genetics* **121**, 357–367.

Zhu, X. & Elston, R. C. (2001). Transmission/Disquelibrium tests for quantitative traits. *Genetic Epidemiology* **20**, 57–74.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **67**, *Part*2 301–320.